Modeling Perceived Quality for Imaging Applications

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op 22 juni 2011 om 12:30 uur
door

Hantao LIU

Master of Science in Signal Processing and Communications
The University of Edinburgh, United Kingdom
geboren te Chengdu, China.

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. I.E.J.R. Heynderickx

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. I.E.J.R. Heynderickx,	Technische Universiteit Delft, promotor
Prof. dr. M. A. Neerincx,	Technische Universiteit Delft
Prof. dr. J. J. Koenderink,	Technische Universiteit Delft
Prof. dr. ir. G. De Haan,	Technische Universiteit Eindhoven
Prof. dr. P. Le Callet,	University of Nantes
Prof. dr. R. Zunino,	University of Genoa

ISBN 978-94-91211-72-0

To my dear parents,
& to my loving wife

Preface

Looking back the time of pursuing a PhD at the Delft University of Technology, I am deeply grateful for all that I have received throughout these years. It has been such a great and unforgettable journey in my life, full of challenge, pleasure, and enlightenment.

First and always, I want to express my most sincere gratitude to my supervisor, Prof. Ingrid Heynderickx. I appreciate all her time and support, both intellectual and financial, which have made this thesis possible and my PhD experience productive, and more importantly, very enjoyable. In particular, I must thank her for giving me the opportunity and freedom to explore, and to pursue my personal research interests. And I must thank her for shaping me into a mature researcher, by developing my scientific attitudes, independent research abilities, and writing and presentation skills. Her encouragement and thoughtful guidance have been the most precious gifts to my academic career. I am also thankful for the excellent example she has provided as a successful professor, not only with her professional expertise, but also with her charming personality.

This thesis would not have been accomplished without the love and encouragement of my family. From the bottom of my heart, I am most grateful to my dear parents for raising me with their unfailing love and always supporting me in all my pursuits. All these years of being abroad, I would not have survived without having them in my life. “Mom and dad, I love you!” Most of all, I would like to thank my wife, Yingying Wu, for her endless love, which truly makes a difference in my life. Her patient, understanding and faithful support during the years of my PhD is so appreciated. “I love you, my darling!”

A very special “thanks!” goes out to Prof. Moncef Gabbouj, Prof. Serkan Kiranyaz, and Prof. Irek Defee from the Tampere University of Technology, where I spent nine months doing research before I started in Delft. Without their motivation I would not have considered a career in scientific research. I would like to thank Prof. Patrick Le Callet for inspiring me to look at my topic from a different perspective. And I am very grateful to Junle, Ulrich, and Judith for their excellent collaboration and for generously sharing their knowledge and experience in the field of image quality assessment. I owe my warm gratitude to Prof. Leon Rothkrantz and Toos for their support in every possible way at the early stages of my PhD in Delft.

My time in the Netherlands has been made pleasurable due to my best friends Xian and Rui who became a part of my life. I am so thankful for the time spent with them, and for our memorable trips around Europe. My time in Delft was also enriched by my colleagues, I would like to thank Nike and Hani for being the great officemates, thank Zhenke, Yangyang, Yun and Chao for all their generous support, thank Wietske for helping me with Dutch language and culture, thank Tim, Alina, and Chang for great conversations and sport time, and thank many other people for sharing with me these years in Delft.

Contents

Abstract

People of all generations are making more and more use of digital imaging systems in their daily lives. The image content rendered by these digital imaging systems largely differs in perceived quality depending on the system and its applications. To be able to optimize the experience of viewers of this content understanding and modeling perceived image quality is essential. Research on modeling image quality in a full-reference framework – where the original content can be used as a reference – is well established in literature. In many current applications, however, the perceived image quality needs to be modeled in a no-reference framework at real-time. As a consequence, the model needs to quantitatively predict perceived quality of a degraded image without being able to compare it to its original version, and has to achieve this with limited computational complexity in order to enable real-time application. Although human beings effortless judge image quality in a real-time no-reference framework, developing a model to simulate this perception is still an academic challenge partly due to our limited understanding of the human visual system.

This thesis presents some achievements in designing no-reference objective quality metrics, which have the aim to automatically and quantitatively predict perceived image quality. Two different approaches are used. In one approach the perception of some specific image degradations is modeled. This approach is applied to the perception of blockiness and ringing, two degradations typically occurring as a consequence of signal compression. The resulting metrics are based on a two-steps framework: a first step, in which the artifacts are located and a second step, in which the local visibility of the artifact is estimated. Both components include aspects of human vision with which the reliability of the metrics in predicting perceived artifact annoyance is improved, while keeping the computational effort limited. In a second approach the overall perceived quality of images is predicted. An accurate and computationally efficient way to do so exists of combining a simplified feature extraction strategy – resulting in features based on aspects of the artifact specific metrics – with an adaptive neural network. After having trained the overall quality estimation system off-line, the metric can be very easily implemented in real-time devices.

Whether the artifacts in an image attract the viewer's attention also affect the viewer's quality estimation. Hence, in a final study the improvement in quality prediction performance of various metrics by including visual attention is evaluated. In these metrics local quality information is weighted with the attention given locally by the averaged viewer. Results show that when using ground-truth attention obtained from eye-tracking recordings the degree to which the quality estimation is improved, depends on the type of metric and kind of image content.

Samenvatting

Mensen van alle generaties gebruiken steeds meer digitale beeldsystemen in hun dagelijks leven. De waargenomen kwaliteit van de beelden, die door deze digitale systemen weergegeven worden, verschilt sterk afhankelijk van het systeem en de toepassingen. Om de kijkervaring te optimaliseren voor de gebruiker van deze beelden is het nodig om waargenomen beeldkwaliteit te begrijpen en modeleren. Onderzoek naar beeldkwaliteit in een situatie waarbij de originele beelden als een referentie aanwezig zijn is overvloedig aanwezig in de literatuur. Maar in vele toepassingen moet de waargenomen beeldkwaliteit geschat worden zonder de aanwezigheid van de originele beelden, en dat in realtime. Dus moet het bijbehorende model de beeldkwaliteit voorspellen zonder het beeld te kunnen vergelijken met zijn originele versie, en dat met een beperkte rekencapaciteit zodat realtime gebruik mogelijk is. Voor mensen is dit een vanzelfsprekende taak, maar deze vanzelfsprekendheid inbouwen in een systeem is nog steeds een academische uitdaging omwille van onze beperkte kennis van het menselijk visuele systeem.

Deze thesis beschrijft een aantal resultaten in het ontwerp van referentievrije objectieve kwaliteitsmaten, die de waargenomen beeldkwaliteit automatisch en kwantitatief moeten voorspellen. Daarbij zijn twee benaderingen gebruikt. In de ene benadering wordt de waarneming van specifieke artefacten in het beeld gemodelleerd. Deze benadering wordt toegepast bij het modeleren van blokkerigheid en rimpeligheid, twee artefacten die voorkomen als een gevolg van beeldcompressie. De bijbehorende objectieve maten zijn gebaseerd op een raamwerk bestaande uit twee stappen: in een eerste stap wordt de locatie van de artefacten bepaald, en in een tweede stap wordt op die locatie de zichtbaarheid van het artefact geschat. In beide stappen worden aspecten van het menselijke visuele systeem meegenomen, waardoor de betrouwbaarheid van de resulterende maten toeneemt, terwijl de rekencomplexiteit beperkt blijft. In de tweede benadering wordt de totale beeldkwaliteit voorspeld. Een manier om dit te doen, die tegelijkertijd nauwkeurig is, maar toch beperkt in rekencomplexiteit, bestaat uit het combineren van een eenvoudig proces voor het berekenen van kenmerken van de artefacten, gebaseerd op aspecten van de bestaande objectieve maten, met een adaptief neuraal netwerk. Zodra het netwerk voldoende getraind is om totale beeldkwaliteit te kunnen voorspellen, kan het eenvoudig in realtime systemen ingebouwd worden.

Het al dan niet trekken van aandacht van de artefacten in een beeld, zal het oordeel over de beeldkwaliteit beïnvloeden. Om deze gedachte ook toe te passen in objectieve kwaliteitsmaten is een aparte studie uitgevoerd. Daarbij worden de lokale zichtbaarheid van de artefacten gewogen met de aandacht die een gemiddelde gebruiker dat deel van het beeld geeft. De gevonden resultaten laten zien dat wanneer actuele data van oogbewegingen gebruikt worden als maat voor de visuele attentie, de mate waarin de voorspelbaarheid van de objectieve maten toeneemt afhankelijk is van de specifieke maat en de beeldinhoud.

Chapter 1

Introduction

During the past decades we have witnessed a revolutionary growth in the use of digital imaging systems in our daily lives. These systems are now part of a broad range of applications, covering communication, entertainment, medical information representation, and security. The signals addressing these imaging systems originate from a diversity of sources. Image content can be either computer generated, or can be recorded with a simple image camera or with a more sophisticated video capturing device. The resulting content is either stored on a memory device or is transmitted over a (broadcasting) channel. In the digital image system the signal is usually processed (e.g. to scale the incoming content to an appropriate spatial and temporal resolution) before being rendered on a display (or a print). Each of these phases in the digital imaging chain is prone to signal distortions [1]-[4]. For example, acquisition of the content is limited in spatial and temporal resolution and its luminance and color information is written in a format with a limited number of bits. For storing or transmitting the content limited bandwidth is available, and as a consequence, the signal is often compressed resulting in a loss of information. In the final step of the chain, the content is rendered on a display or a print with a quality determined by the characteristics of the rendering device. In summary, the quality of the final content is rarely perfect. As a consequence, being able to maintain, control, or even enhance the quality of images has emerged as one of the crucial aspects in the design of current imaging systems [5].

Since human beings are the ultimate receivers of most visual information, image quality has been traditionally evaluated by human observers. When conducted properly, subjective experiments that require the participation of a number of human observers are considered so far as the most reliable means of assessing image quality. However, performing subjective experiments is time-consuming, and therefore, very expensive, and too slow to be useful in real-time applications. To make image quality assessment more applicable, in the last decades, a large amount of research effort has been devoted to the development of computational models/ algorithms that can automatically and consistently predict perceived image quality (see e.g. [5]-[7]). The essential goal is to emulate or at least come close to human perception of image quality, using today's computational technologies. These image quality assessment models/ algorithms are generally referred to as *objective metrics*.

Objective metrics, nowadays, are taking an increasingly important role in a wide variety of applications [5]. Firstly, they can be used in a conventional quality control system to monitor and adjust the image quality in real-time of e.g. a videoconferencing system. A network provider can use a quality metric to examine the video quality transmitted over the network, and to dynamically allocate streaming resources accordingly. Secondly, they can be employed off-line to evaluate image and video processing algorithms. New successful objective metrics are replacing the mean squared error (MSE) or peak signal-to-noise ratio (PSNR) for

comparing the quality performance of competing image and video processing algorithms [8]. In addition, these more sophisticated metrics can be used to optimize the settings of image enhancement algorithms towards the best perceived quality. Thirdly, objective metrics can be embedded in real-time digital imaging systems to optimize their performance from the human perception point of view. If sufficiently in agreement with human quality judgment at a complexity that allows implementation in a real-time imaging chain, the objective metrics are used to optimize the output of the chain. More particularly, they can be adopted to adjust the parameter settings of all video processing algorithms in the chain, taking into account characteristics of the incoming signal. In conclusion, the tremendous demand for image and video technologies has boosted the requirement for a reliable assessment of their quality perceived by users, and has given the field of objective metric development a lot of attention in the past years.

In literature [5]-[7], objective metrics are generally classified into full-reference (FR), reduced-reference (RR) and no-reference (NR) metrics, depending on to what extent they use the original, distortion-free image as a reference. FR metrics assume that the reference is fully accessible, and they are based on measuring the similarity or fidelity between the distorted image and its original version. RR metrics are mainly used in scenarios where the reference is not fully available, e.g. in complex communication networks. They make use of certain features extracted from the reference, which are then employed as side information to evaluate the quality of a distorted image. In many real-time applications, however, there is no access to the reference at all. Hence, it is desirable to have NR metrics that can assess perceived quality based on the distorted image only. Designing NR metrics has great potential in practice, but is still an academic challenge partly due to our limited understanding of the human visual system (HVS).

1.1 Subjective Image Quality Assessment

Subjective image quality is of fundamental importance to the design and validation of objective metrics [7], [9]. It provides a better understanding of how quality is assessed by the HVS, and this understanding greatly helps for mapping objective quality prediction to subjective quality experience. In addition, quality scores resulting from subjective experiments are widely accepted as the benchmark for evaluating the performance of an objective metric, and for comparing alternative metrics proposed in the literature. To obtain useful and reliable results from subjective experiments, it is necessary to design an experimental protocol that best fits the goal of the image quality assessment problem at hand [9], [10]. In this protocol aspects related to viewing conditions, test material and test methods have to be discussed and selected. Typical issues regarding subjective tests are documented in [7] and [9]. In recent years, the dramatic increase in research on objective quality metrics has pushed the need for public, freely available, databases of images/ videos and their corresponding subjective quality scores to the forefront. Having these databases largely facilitates the development of new objective metrics and their performance evaluation in a comparative setting with existing metrics. A direct comparison on the same content and quality scores allows an analysis of the strengths and weaknesses of all metrics available. Some of the databases are summarized in [11].

1.2 Objective Metrics: from MSE/PSNR to State-of-the-art

The most well-known and widely used objective metric is MSE/ PSNR. It is a FR metric that simply sums all pixel-by-pixel differences between a distorted image and its original version. The metric is parameter free, and very inexpensive to implement, but it is widely criticized by the image quality community for its poor correlation with human perceived image quality [8]. MSE/ PSNR, ironically, remains the most used quality metric in current signal and image processing systems, mainly because it is a convention.

Researchers have taken different approaches to develop FR metrics with a better performance than MSE/ PSNR; mainly by including aspects of the HVS. In order to be able to do so, functional aspects of the HVS needed to be modeled. Advances in human vision research increased our understanding of the structural and functional mechanisms of the HVS, and allowed expressing these psychophysical findings into mathematical models [12]-[15]. Although these models still remain limited in their sophistication, and thus also in their reliability, they are already of great interest to explore their added value in image quality research. One way to integrate HVS aspects in the design of an objective quality metric is defined rather "bottom-up" and simulates well-known functionalities of the early HVS [12]. These metrics, of which numerous different implementations are discussed in literature, are based on a so-called error-visibility framework [5]. This framework decomposes the image signal into channels of various frequencies and orientations in order to reflect human vision at the neural cell level. Classical HVS models, such as the contrast sensitivity function (CSF) per channel, and interactions between these channels to simulate masking, are then implemented. Pioneering work on this approach is described in [16], and more representative models that are consistent with the error-visibility framework are summarized in [5]. Although well studied, there still are some limitations to these metrics. First, our knowledge of the HVS is far from complete, and simulating precisely all related components of the HVS is impossible. This intrinsically limits the accuracy of these metrics. Second, the HVS is a rather complex system that contains many nonlinear operations. But, most existing vision models are linear (or quasi-linear) and are developed using restricted and simplistic stimuli. Applying these vision models in objective metric design actually implies the acceptance of a number of strong assumptions.

There are two recent and very successful alternatives to achieve a reliable FR metric. They both are based on a higher level "top-down" approach of the overall functionality of the HVS. It concerns the "structural similarity" (SSIM) [17] and the "visual information fidelity" (VIF) [18] metrics. The principal idea behind SSIM is the observation that the HVS is highly adapted to extract structural information from visual scenes. Therefore, the metric intends to quantify image quality by measuring the structural similarity (or distortion) between a distorted image and its original version. SSIM defines nonstructural distortions as those that do not modify the structure of objects in the visual scene, whereas all other distortions are defined as structural. The metric measures the similarity in three elements, i.e. luminance, contrast, and structure, within a local area of image content. The design of VIF is based on an information communication and sharing point of view. It attempts to relate image quality to the amount of information that is shared between the distorted image and its original version. In other words, VIF exploits the

relationship between statistical image information and image quality. It has been shown that both SSIM and VIF are much more consistent than MSE/ PSNR in predicting perceived image quality. A comprehensive evaluation of the performance of SSIM, VIF as well as other recent FR metrics is detailed in [19].

Compared to the research on FR metrics, that on NR metrics is still in a very preliminary stage. Nonetheless, research on NR metrics has recently received a lot of attention, because of their great practical potential in real-time applications. Assessing quality based on the distorted image only seems an easy task for human observers, yet it is the most difficult problem in objective image quality metric design [5]. Fortunately, in many practical applications, the processes involved in generating the distortions are known and fixed, and so, the design of a NR metric that handles a specific distortion type turns out to be much more realistic. Based on this idea, NR metrics can be categorized into general metrics and dedicated metrics. General NR metrics are intended to assess the overall perceived quality of an image degraded by a known distortion process, which possibly contains various artifact types, e.g. a wavelet-based compressed image often exhibits blur and ringing artifacts simultaneously [20]. Dedicated NR metrics instead are based on directly measuring a specific artifact type created by a specific image distortion process, such as blur caused by image acquisition or blocking artifacts resulting from block-based DCT coding [20].

In the design of a general NR metric, the overall quality of specifically degraded images is often targeted using hypothesized assumptions about natural scenes or the HVS. The NR approach proposed in [21] relies on the assumption that images of natural scenes exhibit strong statistical regularities, and therefore, reside in a tiny area of the space containing all possible images. As a consequence, it quantifies the overall quality of images compressed by JPEG2000 based on detected variations in the statistics of image features calculated in the wavelet domain. The performance of this approach, however, largely depends on sophisticated modeling of natural scene statistics. As an alternative, NR image quality assessment is formulated as a machine learning problem in some research (such as e.g. in [22]-[24]). It avoids the explicit modeling of the HVS, but rather treats it as a black box, whose input-output relationship between image characteristics and a quality rating is to be learned by computational intelligent tools, such as neural networks. This type of NR metrics is generally defined as a regression or function approximation, and therefore, usually requires extensive training on a large data set obtained from subjective quality rating experiments. The two types of general NR metrics mentioned so far have been proved to be effective for the overall quality prediction of a specific combination of distortions, but they are unlikely to be able to handle other combinations of distortions. For example, a NR metric based on a neural network that is trained to assess the quality of JPEG compressed images is not necessarily useful to predict the perceived quality of JPEG2000 compressed images.

In the literature, a large number of NR metrics are designed to assess the quality degradation of a specific type of artifact, such as blockiness, ringing, or noise. These dedicated NR metrics are highly beneficial for image/ video compression and transmission systems. First, they usually provide a spatially varying quality degradation profile of a distorted image, indicating at each location in the image the visibility of the targeted type of artifact. Second, these metrics can each individually determine the quality degradation caused by a specific type of artifact, e.g. the

annoyance of blockiness and ringing can be quantified simultaneously and separately for each JPEG compressed image. Both aspects contribute to the optimization of signal enhancement at either local or global level in an imaging chain. For example, in the video chain of current television sets, the artifact reduction scheme uses these metrics to quantify the occurrence of individual artifacts in the incoming video, and automatically adjusts the algorithms and their parameter settings accordingly [25]-[28]. Finally, the overall image quality, when needed, can be predicted by combining individual artifact specific metrics; e.g. a ringing metric and a blur metric are often combined to assess the overall perceived quality of wavelet-based compressed images [29]. So, both application scenarios illustrate the added value of reliably modeling specific types of artifacts.

In some specific application environments, especially in multimedia communication, RR metrics are used as a compromise between FR and NR metrics [5], [30]. In the context of communication, identification of the quality loss in the video data transmitted over complex networks is highly needed. In such a scenario, FR metrics cannot be applied since there is no access to the original video data at the receiving side. On the other hand, NR metrics have limited reliability since the type of distortions occurring in complex communication networks can be insufficiently predicted. A RR approach provides a practical solution; it only sends partial information about the reference as additional data from the transmitter to the receiver. Obviously, the bandwidth needed for sending the additional information becomes a crucial aspect in the metric design.

So far, advances in image quality assessment have shown the need and practical attainability of integrating relevant aspects of the HVS in objective metric design. In the literature, lower level aspects of the HVS, such as contrast sensitivity, luminance masking and texture masking, are successfully modeled and integrated in various metrics. Studies evaluating whether also higher level aspects of the HVS, such as visual attention, are beneficial for objective quality prediction, and if so, how to apply them in metric design are still limited, but recently have emerged as an active research area [30]-[34]. Adding visual attention in an objective metric is not a trivial task due to the fact that the mechanisms of attention for image quality judgment are not fully understood yet. Until recently, very little, if any, meaningful progress has been made in this scientific direction.

1.3 Aim

As discussed above, designing NR metrics that reliably predict what humans perceive is still challenging, and only limited progress has been made in this research area. Nonetheless, the research area is highly important because of its practical use in improving the quality performance of digital imaging systems. The aim of this thesis is to contribute new developments in the design of NR image quality metrics that quantify the perceived annoyance of specific visual distortions. More specifically, the research described in this thesis provides additional evidence to the research questions:

- what is the added value of adding HVS characteristics to the design of specific NR metrics?
- what is the added value of adding visual attention to the design of objective metrics?

1.4 Outline of the Thesis

One of the difficulties in the design of specific NR metrics is that most types of artifacts are image content and/ or application dependent. As such, the task of precisely locating these artifacts is difficult, especially in a NR context. Another challenge is that the visibility of these artifacts to the human eye is often affected by local image characteristics. Measuring visibility largely relies on modeling the HVS, which involves understanding the way human beings perceive a specific artifact type, and modeling that perception in a computationally efficient way, the latter becoming extremely important for real-time implementation. In chapter 2, 3 and 4 we focus our research to the development of a NR metric for blocking and ringing artifacts, which typically occur in current image/ video compression and transmission. To model the perception of these artifacts, we propose an approach which intrinsically exists of two steps: first detecting regions in an image where artifacts might occur, and second quantifying the artifact annoyance in these regions. In both steps, the specific physical structure of the targeted artifact and properties of the HVS are efficiently combined to characterize the visibility of artifacts to the human eye.

Most existing blockiness metrics are implemented as predictor of overall image quality degradation due to DCT coding; they do not consider local visibility of blocking artifacts. This implies that these metrics do not give precise information on how annoying blocking artifacts at a local level are. To overcome this issue we explicitly introduce in chapter 2 two essential components in the metric design: (1) the detection of the exact location of blocking artifacts independent of e.g. deviations due to spatial scaling of the image, and (2) the estimation of the local visibility of blocking artifacts, based on modeling spatial masking properties of the HVS. These extensions with respect to existing metrics serve two purposes. First, our approach intrinsically yields a spatially varying degradation profile, which is beneficial for applications, where image content can be processed locally adaptive. Second, the overall perceived blockiness is more reliably predicted by only summing the local contributions in the image, where blockiness is perceived.

Unlike the blocking artifact, perceived ringing is rather difficult to be predicted and modeled computationally due to its strong image content dependency. To the best of our knowledge, only a very limited amount of research has been devoted to the development of a ringing metric. To better understand how human beings perceive ringing in compressed images, we conducted two perception experiments: the so-called ringing region visibility experiment (reported in chapter 3) and ringing annoyance experiment (reported in chapter 4). For the ringing region visibility experiment, participants were requested to mark any region in the image where ringing was perceived, independent of its annoyance. For the ringing annoyance experiment, participants scored the annoyance of the ringing artifacts. The resulting subjective data are used in the design of a ringing metric that follows a similar two-step approach as used for our blockiness metric. First, perceptually relevant ringing artifacts are detected, using a perceptual edge detector combined with an efficient model of spatial masking in the HVS (as detailed in chapter 3). Then, the supra-threshold visibility of ringing artifacts within the detected regions is estimated, and the overall ringing annoyance in an image is predicted (as detailed in chapter 4).

Reliably assessing overall quality of images, in which various types of artifacts are coexisting, is still challenging in a NR context. Combining dedicated NR metrics to an overall perceived quality prediction is promising, and consequently, a more complex system for overall quality prediction constructed by several (relevant) artifact specific metrics (including, for example, the blockiness metric and ringing metric developed in this thesis) is highly expected. This approach, however, is so far limited by the inadequate progress in the design of all artifact-specific metrics needed, and by our insufficient understanding of how humans combine various perceived artifacts to an overall quality judgment. To have a more practical solution, an alternative NR approach for the overall image quality assessment is proposed in chapter 5. The basic idea is to efficiently select and calculate the most relevant feature(s) representative for the overall image quality, and to apply an adaptive neural network to empirically learn the highly nonlinear relationship between the relevant feature(s) and the overall image quality assessment. We have shown that skillfully combining the simplified feature computation with the neural network processing yields indeed a promising NR metric for assessing the overall quality of JPEG/ JPEG2000 compressed images. The features selected as input to the neural network are based on local blockiness for JPEG and local blur for JPEG2000, respectively. In a neural network approach these features seem to be sufficiently representative for overall image quality. We could also have chosen a combination of features, e.g. based on local blockiness and ringing for JPEG, or local blur and ringing for JPEG2000, but we decided to limit the input space of the neural network to a single type of features, since these are the simplest to be calculated. As such, the proposed approach is simple, computationally inexpensive, and can be easily implemented in real-time applications.

Novel 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 distortions with local saliency, a process referred to as “visual importance pooling”. 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 use “ground truth” visual attention data instead of a computational model, thus making the evaluation of adding visual attention in objective metrics independent of the reliability of an attention model. These “ground truth” visual attention data were obtained from eye-tracking experiments detailed in chapter 6. We performed two eye-tracking experiments: one in which the participants looked freely to undistorted images, and a second one, in which different participants were asked to score the quality of a JPEG compressed version of the images. We intend to answer two questions. First, what is the difference in human attention between free-looking and image quality assessment? And second, what type of visual attention – if any – should be included in objective metrics? Based on our eye-tracking data, we further evaluate their influence on the performance of several objective metrics well-known in literature.

Finally, in chapter 7 we discuss the findings in this thesis in more general terms and give some conclusions on the added value of including HVS characteristics and visual attention in the design of objective metrics. In addition, we give some suggestions for future research in this area. It might be relevant to realize that

chapters 2, 3, 4, 5 and 6 are based on selected publications of the author. Therefore, overlapping information in the introductory section between chapters may be found. However, to maintain consistency in each individual chapter, the original introduction is given.

1.5 References

[1] R. C. Gonzalez and R. E. Woods, Digital Image Processing (2nd Edition), Prentice Hall, 2002.

[2] H. J. Trussell and M. J. Vrhel, Fundamentals of Digital Imaging. Cambridge University Press, 2008.

[3] G. de Haan, Video Processing for Multimedia Systems. Eindhoven, 2000.

[4] R. Lagendijk and J. Biemond, Basic Methods for Image Restoration and Identification, Handbook of Image and Video Processing 2nd edition. Elsevier Academic Press, 2005.

[5] Z. Wang and A. C. Bovik, Modern Image Quality Assessment. Synthesis Lectures on Image, Video, & Multimedia Processing, Morgan & Claypool, San Refael, Calif, USA, 2006.

[6] Stefan Winkler, Digital Video Quality. John Wiley & Sons, 2005.

[7] VQEG: Final report from the video quality experts group on the validation of objective models of video quality assessment. Available: http:/ / www.vqeg.org.

[8] Z. Wang and A. C. Bovik, "Mean Squared Error: Love it or Leave it?" IEEE Signal Processing Magazine, Jan, 2009.

[9] ITU-R Recommendation BT.500-11, Methodology for the subjective assessment of the quality of television pictures, International Telecommunication Union, Geneva, Switzerland, 2002.

[10] H. de Ridder, "Cognitive issues in image quality measurement," Journal of Electronic Imaging, 10(1): 47-55, 2001.

[11] S. Winkler, "Image and Video Quality Resources," http:/ / stefan.winkler.net/ resources.html.

[12] A. B. Watson, Digital Images and Human Vision. The MIT Press, Cambridge, MA, 1993.

[13] B. A. Wandell, Foundations of Vision. Sinauer Assiciates, Inc., 1995.

[14] W. S. Geisler and M. S. Banks, Visual performance. In M. Bass (Ed.), Handbook of Optics. McGraw-Hill, 1995.

[15] L. K. Cormack, Computational models of early human vision. In A. C. Bovik (Ed.), Handbook of Image and Video Processing, 2nd ed. Elsevier Academic Press, April 2005.

[16] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Information Theory*, 4:525-536, 1974.

[17] 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, April 2004.

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

[19] H. R. Sheikh, M. F. Sabir and A. C. Bovik, “A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms,” IEEE Trans. Image Processing, vol. 15, no. 11, pp. 3440-3451, 2006.
[20] M. Yuen and H. R. Wu, “A survey of hybrid MC/ DPCM/ DCT video coding distortions,” *Signal Processing*, vol. 70, no. 3, pp. 247-278, November 1998.
[21] H. R. Sheikh, A. C. Bovik, and L. K. Cormack, "No-Reference Quality Assessment Using Natural Scene Statistics: JPEG2000," IEEE Transactions on Image Processing, vol. 14, no. 11, pp. 1918-1927, December 2005.
[22] P. Le Callet, C. Viard-Gaudin and D. Barba, “A convolutional neural network approach for objective video quality assessment,” IEEE Transactions on Neural Networks, vol. 17, no. 5, pp. 1316-1327, 2006.
[23] P. Gastaldo and R. Zunino, “Neural networks for the no-reference assessment of perceived quality,” Journal of. Electronic Imaging, 14 (3), 033004, 2005.
[24] R. V. Babu, S. Suresh and A. Perkis, “No-reference JPEG-image quality assessment using GAP-RBF,” Signal Processing, vol. 87, no.6, pp.1493-1503, 2007.
[25] S. Cvetkovic, J. Schirris, P. de With, “Non-Linear Locally-Adaptive Video Contrast Enhancement Algorithm Without Artifacts,” IEEE Transactions on Consumer Electronics, 2008.
[26] I. O. Kirenko, R. Muijs, and L. Shao, “Coding artifact reduction using nonreference block grid visibility measure,” in Proc. IEEE Int. Conf. Multimedia Expo, Jul. 2006, pp. 469–472.
[27] L. Shao, J. Wang, I. Kirenko and G. de Haan, "Quality adaptive trained filters for compression artifacts removal," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 897-900, 2008.
[28] J. G. Puttenstein, I. Heynderickx and G. de Haan, “Evaluation of objective quality measures for noise reduction in TV-systems,” Signal Processing: Image Communication, vol. 19, no. 2, pp. 109-119, 2004.
[29] P. Marziliano, F. Dufax, S. Winkler and T. Ebrahimi, “Perceptual blur and ringing metrics: Application to JPEG2000,” Signal Processing: Image Communication, vol. 19, pp. 163-172, 2004.
[30] M. Carnec, P. Le Callet and D. Barba, “Objective quality assessment of color images based on a generic perceptual reduced reference,” Signal Processing: Image Communication, vol. 23 , no. 4, pp. 239-256, 2008.
[31] A. K. Moorthy and A. C. Bovik, “Visual Importance Pooling for Image Quality Assessment,” IEEE Journal of Selected Topics in Signal Processing, Special Issue on Visual Media Quality Assessment, vol. 3, no.2, April 2009.
[32] H. Liu and I. Heynderickx, “Studying the Added Value of Visual Attention in Objective Image Quality Metrics Based on Eye Movement Data,” in *Proc. ICIP*, 2009.
[33] O. Le Meur, A. Ninassi, P. Le Callet and D. Barba, “Overt visual attention for free-viewing and quality assessment tasks,” *Elsevier, Signal Processing: Image Communication*, 2010.
[34] O. Le Meur, A. Ninassi, P. Le Callet and D. Barba, “Do video coding impairments disturb the visual attention deployment?” Signal Processing: Image Communication, vol. 2, no. 8, 2010.

Chapter 2

A Perceptually Relevant No-Reference Blockiness Metric Based on Local Image Characteristics

***Abstract*: A novel no-reference blockiness metric that provides a quantitative measure of blocking annoyance in block-based DCT coding is presented. The metric incorporates properties of the human visual system (HVS) to improve its reliability, while the additional cost introduced by the HVS is minimized to ensure its use for real-time processing. This is mainly achieved by calculating the local pixel-based distortion of the artifact itself, combined with its local visibility by means of a simplified model of visual masking. The overall computation efficiency and metric accuracy is further improved by including a grid detector to identify the exact location of blocking artifacts in a given image. The metric calculated only at the detected blocking artifacts is averaged over all blocking artifacts in the image to yield an overall blockiness score. The performance of this metric is compared to existing alternatives in literature and shows to be highly consistent with subjective data at a reduced computational load. As such, the proposed blockiness metric is promising in terms of both computational efficiency and practical reliability for real-life applications.**

This chapter is based on the research article published as "A Perceptually Relevant No-Reference Blockiness Metric Based on Local Image Characteristics" by H. Liu and I. Heynderickx in EURASIP Journal on Advances in Signal Processing, vol. 2009, 2009.

2.1 Introduction

Objective metrics, which serve as computational alternatives for expensive image quality assessment by human subjects, are aimed at predicting perceived image quality aspects automatically and quantitatively. They are of fundamental importance to a broad range of image and video processing applications, such as for the optimization of video coding or for real-time quality monitoring and control in displays [1], [2]. For example in the video chain of current TV-sets, various objective metrics, which determine the quality of the incoming signal in terms of blockiness, ringing, blur, etc. and adapt the parameters in the video enhancement algorithms accordingly, are implemented to enable an improved overall perceived quality for the viewer.

In the last decades, a considerable amount of research has been carried out on developing objective image quality metrics, which can be generally classified into two categories: full-reference (FR) metrics and no-reference (NR) metrics [1]. The FR metrics are based on measuring the similarity or fidelity between the distorted image and its original version, which is considered as a distortion-free reference. However, in real-world applications the reference is not always fully available; for example, the receiving end of a digital video chain usually has no access to the original image. Hence, objective metrics used in these types of applications are constrained to a no-reference approach, which means that the quality assessment relies on the reconstructed image only. Although, human observers can easily judge image quality without any reference, designing NR metrics is still an academic challenge mainly due to the limited understanding of the human visual system [1]. Nevertheless, since the structure information of various image distortions is well known, NR metrics designed for specific quality aspects rather than for overall image quality are simpler, and therefore, more realistic [2].

Since the human visual system (HVS) is the ultimate assessor of most visual information, taking into account the way human beings perceive quality aspects, while removing perceptual redundancies, can be greatly beneficial for matching objective quality prediction to human perceived quality [5]. This statement is adequately supported by the observed shortcoming of the purely pixel-based metrics, such as the mean square error (MSE) and peak signal-to-noise ratio (PSNR). They insufficiently reflect distortion annoyance to the human eye, and thus often exhibit a poor correlation with subjective test results (e.g. in [1]). The performance of these metrics has been enhanced by incorporating certain properties of the HVS (e.g. in [13]-[16]). But since the HVS is extremely complex, an objective metric based on a model of the HVS often is computationally very intensive. Hence, to ensure that an HVS based objective metric is applicable to real-time processing, investigations should be carried out to reduce the complexity of the HVS model as well as of the metric itself without significantly compromising the overall performance.

One of the image quality distortions for which several objective metrics have been developed is blockiness. A blocking artifact manifests itself as an artificial discontinuity in the image content, and is known to be the most annoying distortion at low bit-rate DCT coding [24]. Most objective quality metrics either require a reference image or video (e.g. in [14]-[16]), which restricts their use in real-life applications, or lack an explicit human vision model (e.g. in [25], [26]), which limits

their reliability. Apart from these metrics, no-reference blockiness metrics, including certain properties of the HVS are developed. Recently, a promising approach, which we refer to as feature extraction method, is proposed in [6] and [7], where the basic idea is to extract certain image features related to the blocking artifact and to combine them in a quality prediction model with the parameters estimated from subjective test data. The stability of this method, however, is uncertain since the model is trained with a limited set of images only, and its reliability to other images is not proved yet.

A no-reference blockiness metric can be formulated either in the spatial domain or in the transform domain. The metrics described e.g. in [8] and [9] are implemented in the transform domain. In [8], a 1-D absolute difference signal is combined with luminance and texture masking, and from that blockiness is estimated as the peaks in the power spectrum using FFT. In this case, the FFT has to be calculated many times for each image, which is therefore very expensive. The algorithm in [9] computes the blockiness as a result of a 2-D step function weighted with a measure of local spatial masking. This metric requires the access to the DCT encoding parameters, which are, however, not always available in practical applications.

In this paper, we rely on the spatial domain approach. The generalized block-edge impairment metric (GBIM) [3] is the most well-known metric in this domain. GBIM expresses blockiness as the inter-pixel difference across block boundaries scaled with a weighting function, which simply measures the perceptual significance of the difference due to local spatial masking of the HVS. The total amount of blockiness is then normalized by the same measure calculated for all other pixels in an image. The main drawbacks for GBIM are: (1) the inter-pixel difference characterizes the block discontinuity not to the extent that local blockiness is sufficiently reliably predicated; (2) the HVS model includes both luminance masking and texture masking in a single weighting function, and efficient integration of different masking effects is not considered, hence, applying this model in a blockiness metric may fail in assessing demanding images; and (3) the metric is designed such that the human vision model needs to be calculated for every pixel in an image, which is computationally very expensive. A second metric using the spatial domain is based on a locally adaptive algorithm [4], and is hereafter referred to as LABM. It calculates a blockiness metric for each individual coding block in an image, and simultaneously estimates whether the blockiness is strong enough to be visible to the human eye by means of a just-noticeable-distortion (JND) profile. Subsequently, the local metric is averaged over all visible blocks to yield a blockiness score. This metric is promising and potentially more accurate than GBIM. However, it exhibits several drawbacks: (1) the severity of blockiness for individual artifacts might be under- or over-estimated by providing an averaged blockiness value for all artifacts within this block; (2) calculating an accurate JND profile which provides a visibility threshold of a distortion due to masking is complex, and it cannot predict perceived annoyance above threshold; and (3) the metric needs to estimate the JND for every pixel in an image, which largely increases the computational cost.

Calculating the blockiness metric only at the expected block edges, and not at all pixels in an image strongly reduces the computational power, especially when a complex HVS is involved. To ensure that the metric is calculated at the exact

position of the block boundaries a grid detector is needed since in practice deviations in the blocking grid might occur in the incoming signal, e.g. as a consequence of spatial scaling [10], [11], [25]. Without this detection phase, no-reference metrics might turn out to be useless, as blockiness is calculated at wrong pixel positions.

In this paper, a novel algorithm is proposed to quantify blocking annoyance based on its local image characteristics. It combines existing ideas in literature with some new contributions: (1) a refined pixel-based distortion measure for each individual blocking artifact in relation to its direct vicinity; (2) a simplified and more efficient visual masking model to address the local visibility of blocking artifacts to the human eye; and (3) the calculation of the local pixel-based distortion and its visibility on the most relevant stimuli only, which significantly reduces the computational cost. The resulting metric yields a strong correlation with subjective data. The rest of the paper is organized as follows: Section II details the proposed algorithm, Section III provides and discusses the experimental results, and the conclusions are drawn in Section IV.

2.2 Description of the Algorithm

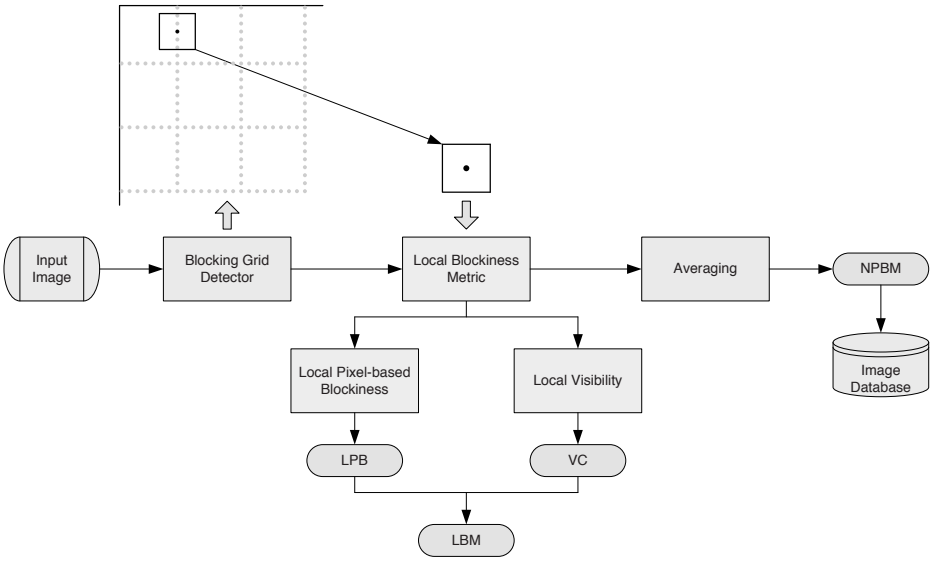


Fig. 1. Schematic overview of the proposed approach.

The schematic overview of the proposed approach is illustrated in Figure 1 (the first outline of the algorithm was already described in [27]). Initially, a grid detector is adopted in order to identify the exact position of the blocking artifacts. After locating the artifacts, local processing is carried out to individually examine each detected blocking artifact by analyzing its surrounding content to a limited extent. This local calculation consists of two parallel steps: (1) measuring the degree of local pixel-based blockiness (LPB); and (2) estimating the local visibility of the artifact to the human eye and outputting a visibility coefficient (VC). The resulting LPB and

VC are integrated into a local blockiness metric (LBM). Finally, the LBM is averaged over the blocking grid of the image to produce an overall score of blockiness assessment (i.e. NPBM). The whole process is calculated on the luminance channel only in order to further reduce the computational load. The algorithm is performed for the blockiness once in horizontal direction (i.e. $NPBM_h$), and once in vertical direction $NPBM_v$. From both values the average is calculated assuming that the human sensitivity to horizontal and vertical blocking artifacts is equal.

2.2.1 Blocking Grid Detection

Since the arbitrary grid problem has emerged as a crucial issue especially for no-reference blockiness metrics, where no prior knowledge on grid variation is available, a grid detector is required in order to ensure a reliable metric [11], [25]. Most, if not all, of the existing blockiness metrics make the strong assumption that the grid exists of blocks of 8x8 pixels, starting exactly at the top-left corner of an image. However, this is not necessarily the case in real-life applications. Every part of a video chain, from acquisition to display, may induce deviations in the signal, and the decoded images are often scaled before being displayed. As a result, grids are shifted, and the block size is changed.

Methods, as e.g. in [8] and [10], employ a frequency-based analysis of the image to detect the location of blocking artifacts. These approaches, due to the additional signal transform involved, are often computationally inefficient. Alternatives in the spatial domain can be found in [11] and [25]. They both map an image into a one-dimensional signal profile. In [11] the block size is estimated using a rather complex maximum-likelihood method, and the grid offset is not considered. In [25] the block size and the grid offset are directly extracted from the peaks in the 1-D signal by calculating the normalized gradient for every pixel in an image. However, spurious peaks in the 1-D signal as a result of edges from objects may occur, and consequently yield possible detection errors. In this paper, we further rely on the basic ideas of both [11] and [25], but implement them by means of a simplified calculation of the 1-D signal and by extracting the block size and the grid offset using DFT of the 1-D signal. The entire procedure is performed once in horizontal and once in vertical direction to address a possible asymmetry in the blocking grid.

1-D Signal Extraction

Since blocking artifacts regularly manifest themselves as spatial discontinuities in an image, their behavior can be effectively revealed through a 1-D signal profile, which is simply formed calculating the gradient along one direction (e.g. horizontal direction), and then summing up the results along the other direction (e.g. vertical direction). We denote the luminance channel of an image signal of MxN (height x width) pixels as $I(i,j)$ for $i \in [1,M], j \in [1,N]$, and calculate the gradient map G_h along the horizontal direction

$$G_h(i,j) = \left| I(i,j+1) - I(i,j) \right|, \quad j \in [1,N-1] \tag{1}$$

The resultant gradient map is reduced to a 1-D signal profile S_h by summing G_h along the vertical direction

$$S_h(j) = \sum_{i=1}^{M} G_h(i, j) \tag{2}$$

Block Size Extraction

Based on the fact that the amount of energy present in the gradient at the borders of coding blocks is greater than that in the intermediate positions, blocking artifacts, if existing, are present as a periodic impulse train of signal peaks. These signal peaks can be further enhanced using some form of spatial filtering, which makes the peaks stand out from their vicinity. In this paper, a median filter is used. Then a promoted 1-D signal profile PS_h is obtained simply subtracting from S_h its median-filtered version MS_h

$$PS_h(j) = S_h(j) - MS_h(j) \tag{3}$$

$$MS_h(j) = Median\ \{S_h(j-k),...,S_h(j),...,S_h(j+k)\} \tag{4}$$

where, the size of the median filter *(2k+1)* depends on *N*. In our experiments, *N* is e.g. 384, and then *k* is 4. The resulting 1-D signal profile PS_h intrinsically reveals the blocking grid as an impulse train with a periodicity determined by the block size. However, in demanding conditions, such as for images with many object edges, the periodicity in the regular impulses might be masked by noise as a result of image content. This potentially makes locating the required peaks and estimating their periodicity more difficult. The periodicity of the impulse train, corresponding to the block size, is more easily extracted from the 1-D signal PS_h in the frequency domain using the Discrete Fourier Transform (DFT).

Grid Offset Extraction

After the block size (i.e. *p*) is determined, the offset of the blocking grid can be directly retrieved from the signal PS_h, in which the peaks are located at multiples of the block size. Thus, a simple approach based on calculating the accumulative value of grid peaks with a possible offset Δx (e.g. $\Delta x = 0:(p-1)$ with the periodic feature in mind), is proposed. For each possible offset value Δx, the accumulator is defined as

$$A(\Delta x) = \sum_{i=1}^{[N/p]-1} PS_h(\Delta x + p \cdot i), \quad \Delta x \in [0, p-1] \tag{5}$$

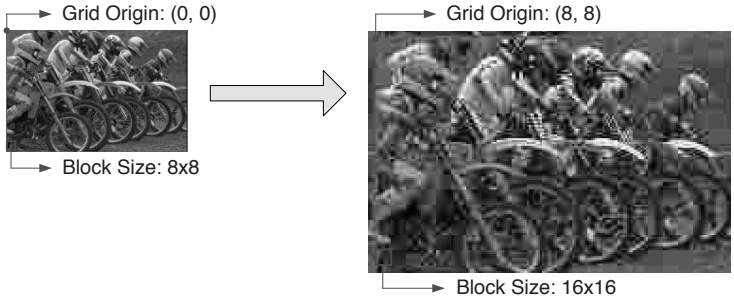
The offset is determined as

$$A(\Delta x) = MAX\ [\ A(0)..A(p-1)\ ] \qquad (6)$$

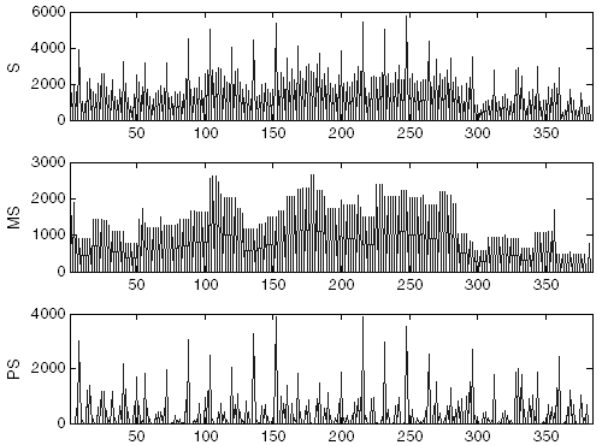
Based on the results of the block size and grid offset, the exact position of blocking artifacts can be explicitly extracted.

An Example

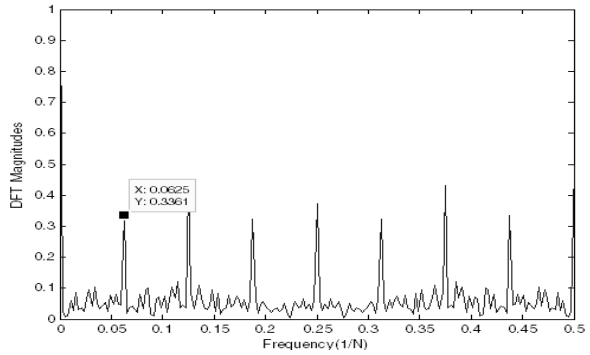
A simple example is given in Figure 2, where the input image "*bikes*" of 128x192 pixels is JPEG-compressed using a standard block size of 8x8 pixels. The displayed image is synthetically up-scaled with a scaling factor 2x2, and shifted by 8 pixels both from left to right and from top to bottom. As a result, the displayed image size is 256x384 pixels, the block size 16x16 pixels, and the grid starts at pixel position (8, 8) instead of at the origin (0, 0), as shown in Figure 2 (a). The proposed algorithm towards a 1-D signal profile is illustrated in Figure 2 (b). Figure 2 (c) shows the magnitude profile of the DFT applied to the signal *PS*. It allows extraction of the period p (i.e. p=1/ 0.0625=16 pixels), which is maintained over the whole frequency range. Based on the detected block size $p = 16$, the grid offset is calculated as $\Delta x = 8$ .Then the blocking grid can be determined, as shown in Figure 2 (d).



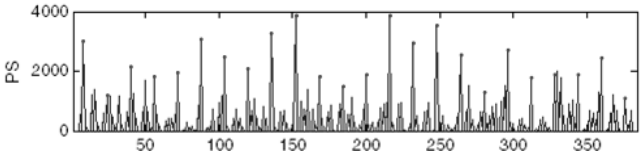
(a) Input image (left) and displayed image (right)



(b) 1-D signal formation: S, MS and PS are calculated according to Equation (2), (3) and (4) for the displayed image in (a) along the horizontal direction



(c) DFT magnitudes of PS in (b)



(d) Blocking grid detected from the displayed image in (a) along the horizontal direction

Fig. 2. Blocking grid detection: an example.

2.2.2 Local Pixel-based Blockiness Measure

Since blocking artifacts intrinsically are a local phenomenon, their behavior can be reasonably described at a local level, indicating the visual strength of a distortion within a local area of image content. Based on the physical structure of blocking artifacts as a spatial discontinuity, this can be simply accomplished relating the energy present in the gradient at the artifact with the energy present in the gradient within its vicinity. This local distortion measure (LDM) purely based on pixel information can be formulated as

$$LDM(k) = \frac{\mathrm{E}^k(i,j)}{f[\mathrm{E}^{V(k)}(i,j)]} \qquad k = 1..n \tag{7}$$

where $f[\cdot]$ indicates the pooling function, e.g. Σ, *mean*, or L2-norm, E^k indicates the gradient energy calculated for each individual artifact and $\mathrm{E}^{V(k)}$ indicates the gradient energy calculated at the pixels in the direct vicinity of this artifact, and n is the total number of blocking artifacts in an image. Since the visual strength of a block discontinuity is primarily affected by its local surroundings of limited extent, this approach is potentially more accurate than a global measure of blockiness (e.g. [3] and [25]), where the overall blockiness is assessed by the ratio of the averaged discontinuities on the blocking grid and the averaged discontinuities in pixels which are not on the blocking grid. Furthermore, the local visibility of a distortion

due to masking can now be easily incorporated, with the result that it is only calculated at the location of the blocking artifacts. This means that modeling the HVS on non-relevant pixels is eliminated as compared to the global approach (e.g. [3]).

In this paper, we rely on the inter-block difference defined in [4], and extend the idea by reducing the dimension of the blockiness measure from a signal block to an individual blocking artifact. As such, the local distortion measure (LDM) is implemented on the gradient map, resulting in local pixel-based blockiness (LPB). The LPB quantifies the blocking artifact at pixel location (i, j) as:

$$LPB_h(i,j) = \begin{cases} \omega \times BG_h & if \quad NBG_h = 0, BG_h \neq 0 \\ \dfrac{BG_h}{NBG_h} & if \quad NBG_h \neq 0 \\ 0 & if \quad NBG_h = 0, BG_h = 0 \end{cases} \tag{8}$$

where BG_h and NBG_h are

$$BG_h = G_h(i,j) \tag{9}$$

$$NBG_h = \frac{1}{2n} \sum_{x=-n..n, x\neq 0} G_h(i, j+x) \tag{10}$$

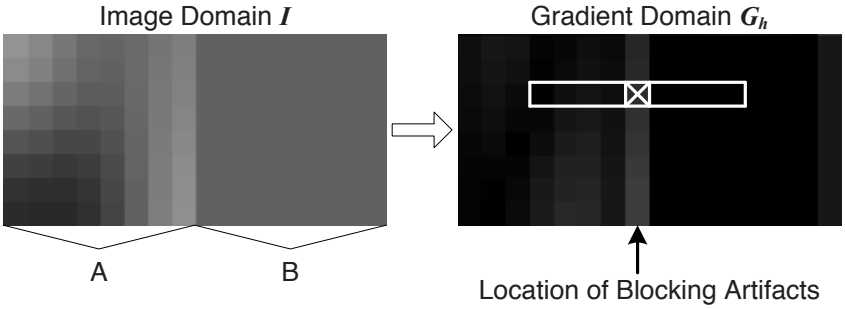


Fig. 3. Local pixel-based blockiness (LPM).

The definition of the LPB is further explained as follows:

(1) The template addressing the direct vicinity is defined as a 1-D element including n adjacent pixels to the left and to the right of an artifact. The size of the template (2n+1) is designed to be proportional to the detected block size p (e.g. n=p/ 2), taking into account possible scaling of the decoded images. An example of the template is shown in Figure 3, where two adjacent 8x8 blocks (i.e. A and B) are extracted from a real JPEG image.

(2) BG_h denotes the local energy present in the gradient at the blocking artifact, and NBG_h denotes the averaged gradient energy over its direct vicinity. If

$NBG_h = 0$, only the value of BG_h determines the local pixel-based blockiness. In this case, $LPB_h = 0$ (i.e. $BG_h = 0$) means there is no block discontinuity appearing, and the blocking artifact is spurious. $LPB_h = \omega \times BG_h$ (i.e. $BG_h \neq 0$) means the artifact exhibits a severe extent of blockiness, and ω ($\omega = 1$ in our experiments) is used to adjust the amount of gradient energy. If $NBG_h \neq 0$, the local pixel-based blockiness is simply calculated as the ratio of BG_h over NBG_h.

(3) The local pixel-based blockiness LPB_h is specified in equations (8) to (10) for a block discontinuity along the horizontal direction. The measure of LPB_v for vertical blockiness can be easily defined in a similar way. The calculation is then performed within a vertical 1-D template.

2.2.3 Local Visibility Estimation

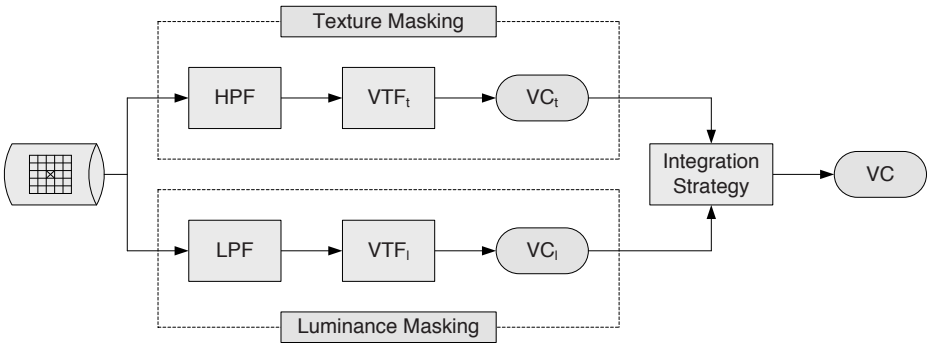


Fig. 4. Schematic overview of the proposed human vision model.

To predict perceived quality, objective metrics based on models of the human visual system are potentially more reliable [5], [19]. However, from a practical point of view, it is highly desirable to reduce the complexity of the HVS model without compromising its abilities. In this paper, a simplified human vision model based on the spatial masking properties of the HVS is proposed. It adopts two fundamental characteristics of the HVS, which affect the visibility of an artifact in the spatial domain: (1) the averaged background luminance surrounding the artifact; and (2) the spatial non-uniformity in the background luminance [18], [19]. They are known as luminance masking and texture masking, respectively, and both are highly relevant to the perception of blocking artifacts.

Various models of visual masking to quantify the visibility of blocking artifacts in images have been proposed in literature [3], [6], [12], [16], [18]. Among these models, there are two widely used ones: the model used in GBIM [3] and the just-noticeable-distortion (JND) profile model used in [18]. Their disadvantages have already been pointed out in Section I. Our proposed model is illustrated in Figure 4. Both texture and luminance masking are implemented by analyzing the local signal properties within a window, representing the local surrounding of a blocking artifact. A visibility coefficient as a consequence of masking (i.e. VC_t and VC_l,

respectively) is calculated using spatial filtering followed by a weighting function. Then, both coefficients are efficiently combined into a single visibility coefficient (VC), which reflects the perceptual significance of the artifact quantitatively.

Local Visibility due to Texture Masking

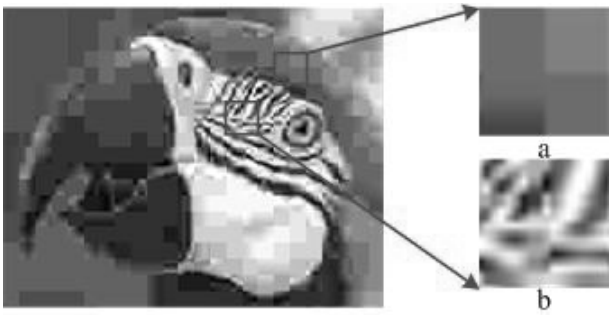


Fig. 5. An example of texture masking on blocking artifacts.

Figure 5 shows an example of texture masking on blocking artifacts, where "a" and "b" are patterns including 4 adjacent blocks of 8x8 pixels extracted from a JPEG-coded image. As can be seen from the right-hand side of Figure 5 pattern "a" and pattern "b" both intrinsically exhibit block discontinuities. However, as shown on the left-hand side of Figure 5, the block discontinuities in pattern "b" are perceptually masked by its non-uniform background, while the block discontinuities in pattern "a" are much more visible as it is in a flat background. Therefore, texture masking can be estimated from the local background activity [19]. In this paper, texture masking is modeled calculating a visibility coefficient (VC_t), indicating the degree of texture masking. The higher the value of this coefficient, the smaller the masking effect, and hence, the stronger the visibility of the artifact is. The procedure of modeling texture masking comprises three steps:

- Texture Detection: calculate the local background activity (non-uniformity).
- Thresholding: a classification scheme to capture the active background regions.
- Visibility Transform Function (VTF): obtain a visibility coefficient (VC_t) based on the HVS characteristics for texture masking.

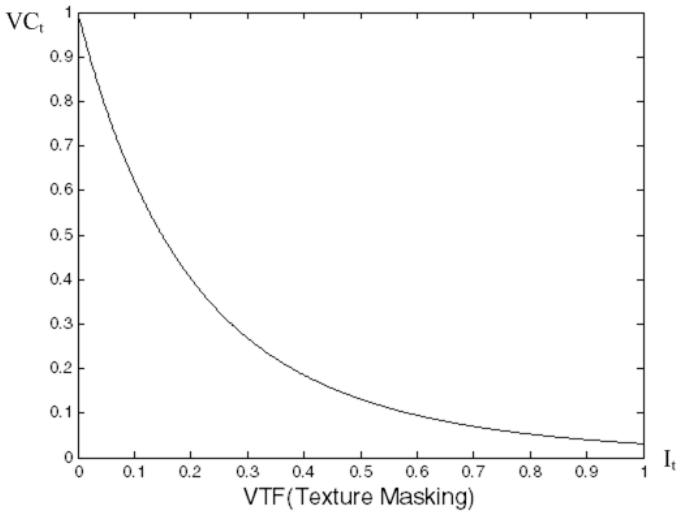
1	2	0	-2	-1
4	8	0	-8	-4
6	12	0	-12	-6
4	8	0	-8	-4
1	2	0	-2	-1

T1

1	4	6	4	1
2	8	12	8	2
0	0	0	0	0
-2	-8	-12	-8	-2
-1	-4	-6	-4	-1

T2

(a) The high-pass filters for texture detection



(b) Visibility transform function (VTF) used

Fig. 6. Implementation of the texture masking.

Texture detection can be performed convolving the signal with some form of high-pass filter. One of the Laws' texture energy filters [20] is employed here in a slightly modified form. As shown in Figure 6, $T1$ and $T2$ are used to measure the background activity in horizontal and vertical direction, respectively. A pre-defined threshold Thr ($Thr = 0.15$ in our experiments) is applied to classify the background into "flat" or "texture", resulting in an activity value $I_t(i,j)$, which is given by

$$I_t(i,j) = \begin{cases} 0 & if\ t(i,j) < Thr \\ t(i,j) & otherwise \end{cases} \tag{11}$$

$$t(i,j) = \frac{1}{48}\sum_{x=1}^{5}\sum_{y=1}^{5} I(i-3+x, j-3+y)\cdot T(x,y) \tag{12}$$

where $I(i,j)$ denotes the pixel intensity at location (i,j), and T is chosen as $T1$ for texture calculation in horizontal direction, and $T2$ in vertical direction. It should be noted that splitting up the calculation in horizontal and vertical direction, and using a modified version of the texture energy filter, in which some template coefficients are removed, can be done having the application of a blockiness metric in mind. The texture filters need to be adopted in case of extending these ideas to other objective metrics.

A visibility transform function (VTF) is proposed in accordance to human perceptual properties, which means that the visibility coefficient $VC_t(i,j)$ is inversely proportional (nonlinear) to the activity value $I_t(i,j)$. Figure 6 shows an example of such a transform function, which can be defined as

$$VC_t(i,j) = \frac{1}{(1+\overline{I}_t(i,j))^{\alpha}} \tag{13}$$

where $VC_t(i,j) = 1$, when the stimulus is in a "flat" background, and $\alpha > 1$ ($\alpha = 5$ in our experiments) is used to adjust the nonlinearity. This shape of the VTF is an approximation, considered to be good enough.

Local Visibility due to Luminance Masking

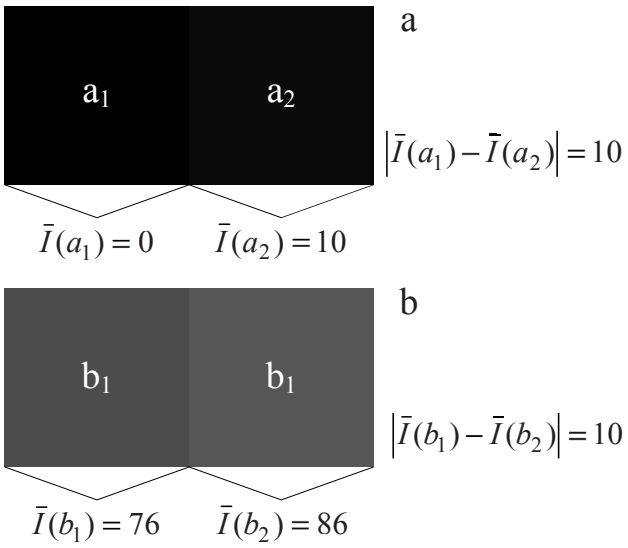


Fig. 7. An example of luminance masking on blocking artifacts.

In many psychovisual experiments it was found that the human visual system's sensitivity to variations in luminance depends on (is a nonlinear function of) the local mean luminance [16], [18], [19], [23]. Figure 7 shows an example of luminance masking on blocking artifacts, where "a" and "b" are synthetic patterns, each of which includes 2 adjacent blocks with different gray scale levels. Although the intensity difference between the two blocks is the same in both patterns, the block discontinuity of pattern "b" is much more visible than that in pattern "a" due to the difference in background luminance. In this paper, luminance masking is modeled based on two empirically driven properties of the HVS: (1) a distortion in a dark surrounding tends to be less visible than one in a bright surrounding [16], [18], and (2) a distortion is most visible for a surrounding with an averaged luminance value between 70 and 90 (centered approximately at 81) in 8bits gray-scale images [23]. The procedure of modeling luminance masking consists of two steps:

- Local Luminance Detection: calculate the local averaged background luminance.
- Visibility Transform Function (VTF): obtain a visibility coefficient (VC_t) based on the HVS characteristics for luminance masking.

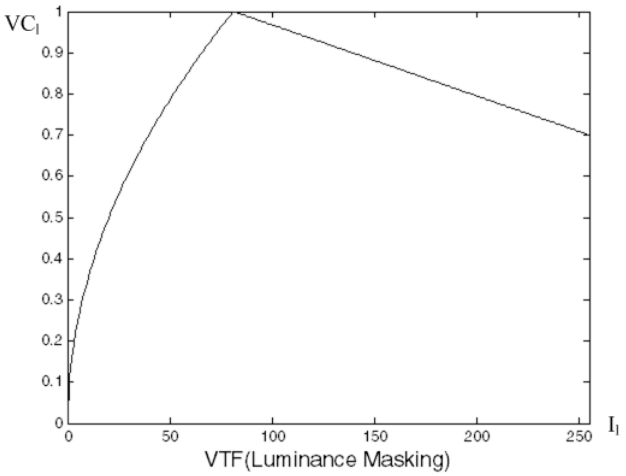
1	1	0	1	1
1	2	0	2	1
1	2	0	2	1
1	2	0	2	1
1	1	0	1	1

L1

1	1	1	1	1
1	2	2	2	1
0	0	0	0	0
1	2	2	2	1
1	1	1	1	1

L2

(a) The low-pass filters for local luminance detection



(b) Visibility transform function (VTF) used

Fig. 8. Implementation of the luminance masking.

The local luminance of a certain stimulus is calculated using a weighted low-pass filter as shown in Figure 8, in which some template coefficients are set to "0". The local luminance $I_l(i,j)$ is given by

$$I_l(i,j) = \frac{1}{26}\sum_{x=1}^{5}\sum_{y=1}^{5} I(i-3+x, j-3+y)\cdot L(x,y) \qquad (14)$$

where L is chosen as $L1$ for calculating the background luminance in horizontal direction, and $L2$ in vertical direction. Again, splitting up the calculation in horizontal and vertical direction, and using a modified low-pass filter, in which some template coefficients are set to 0, is done with the application of a blockiness metric in mind.

For simplicity, the relationship between the visibility coefficient $VC_l(i,j)$ and the local luminance $I_l(i,j)$ is modeled by a nonlinear function (e.g. power law) for low background luminance (i.e. below 81), and is approximated by a linear function at higher background luminance (i.e. above 81). This functional behavior is shown in Figure 8, and mathematically described as

$$VC_l(i,j) = \begin{cases} \left(\frac{I_l(i,j)}{81}\right)^{1/2} & if\ 0 \le I_l(i,j) \le 81 \\ \left(\frac{1-\beta}{174}\right) \cdot (81 - I_l(i,j)) + 1 & otherwise \end{cases} \quad (15)$$

where $VC_l(i,j)$ achieves the highest value of 1 when $I_l(i,j) = 81$, and $0 < \beta < 1$ ($\beta = 0.7$ in our experiments) is used to adjust the slope of the linear part of this function.

Integration Strategy

The visibility of an artifact depends on various masking effects co-existing in the HVS. How to efficiently integrate them is an important issue in obtaining an accurate perceptual model [17]. Since masking intrinsically is a local phenomenon, the locality in the visibility of a distortion due to masking is maintained in the integration strategy of both masking effects. The resulting approach is schematically given in Figure 9. Based on the local image content surrounding a blocking artifact first the texture masking is calculated. In case the local activity in the area is larger than a given threshold (see equation (11)), a visibility coefficient VC_t is applied, followed by the application of a luminance masking coefficient VC_l. In case the local activity in the area is low, only VC_l is applied. The application of VC_l, where appropriately combined with VC_t, results in an output value VC.

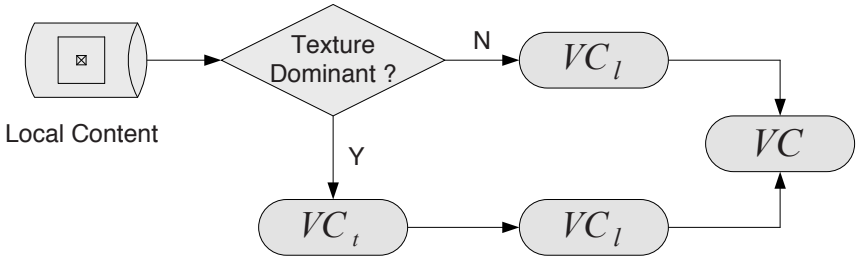


Fig. 9. Integration strategy of the texture and luminance masking effect.

2.2.4 The Perceptual Blockiness Metric

The local pixel-based blockiness (LPB) defined in section II.B is purely signal based, and so does not necessarily yield perceptually consistent results. The human vision model proposed in section II.C aims at removing the perceptually insignificant components due to visual masking. Integration of these two elements can be simply performed at a local level using the output of the human vision model (VC) as a weighting coefficient to scale the local pixel-based blockiness (LPB), resulting in a local perceptual blockiness metric (LPBM). Since the horizontal and vertical blocking artifacts are calculated separately, the LPBM for the block discontinuity along the horizontal direction is described as

$$LPBM_h(i,j) = VC(i,j) \times LPB_h(i,j) \tag{16}$$

which is then averaged over all detected blocking artifacts in the entire image to determine an overall blockiness metric, i.e. a no-reference perceptual blockiness metric (NPBM)

$$NPBM_h = \frac{1}{n}\sum_{k=1}^{n} [LPBM_h(i,j)]_k \tag{17}$$

where n is the total number of pixels on the blocking grid of an image.

A metric $NPBM_v$ can be similarly defined for the blockiness along the vertical direction, and is simply combined with $NPBM_h$ to give the resultant blockiness score for an image. More complex combination laws may be appropriate but need to be further investigated.

$$NPBM = \frac{NPBM_h + NPBM_v}{2} \tag{18}$$

In our case, the human vision model is only calculated at the location of blocking artifact, and not for all pixels in an image. This significantly reduces the computational cost in the formulation of an overall metric.

2.3 Evaluation of the Overall Metric Performance

Subjective ratings resulting from psychovisual experiments are widely accepted as the benchmark for evaluating objective quality metrics. They reveal how well the objective metrics predict the human visual experience, and how to further improve the objective metrics for a more accurate mapping to the subjective data. The LIVE quality assessment database (JPEG) [22] is used to compare the performance of our proposed metric to various alternative blockiness metrics. The LIVE database consists of a set of source images that reflects adequate diversity in image content. Twenty-nine high resolution and high quality color images are compressed using JPEG at a bit rate ranging from 0.15bpp to 3.34bpp, resulting in a database of 233 images. A psychovisual experiment was conducted to assign to each image a mean opinion quality score (MOS) measured on a continuous linear scale that was divided into five intervals marked with the adjectives “Bad”, “Poor”, “Fair”, “Good” and “Excellent”.

The performance of an objective metric can be quantitatively evaluated with respect to its ability to predict subjective quality ratings, based on prediction accuracy, prediction monotonicity, and prediction consistency [21]. Accordingly, the Pearson linear correlation coefficient, the Spearman rank order correlation coefficient, and the outlier ratio are calculated. As suggested in [21], the metric’s performance can also be evaluated with nonlinear correlations using a non-linear mapping function for the objective predictions before computing the correlation. For example, a logistic function may be applied to the objective metric results to account for a possible saturation effect. This way of working usually yields higher

correlation coefficients. Nonlinear correlations, however, have the disadvantage of minimizing performance differences between metrics [12]. Hence, to make a more critical comparison, only linear correlations are calculated in this paper.

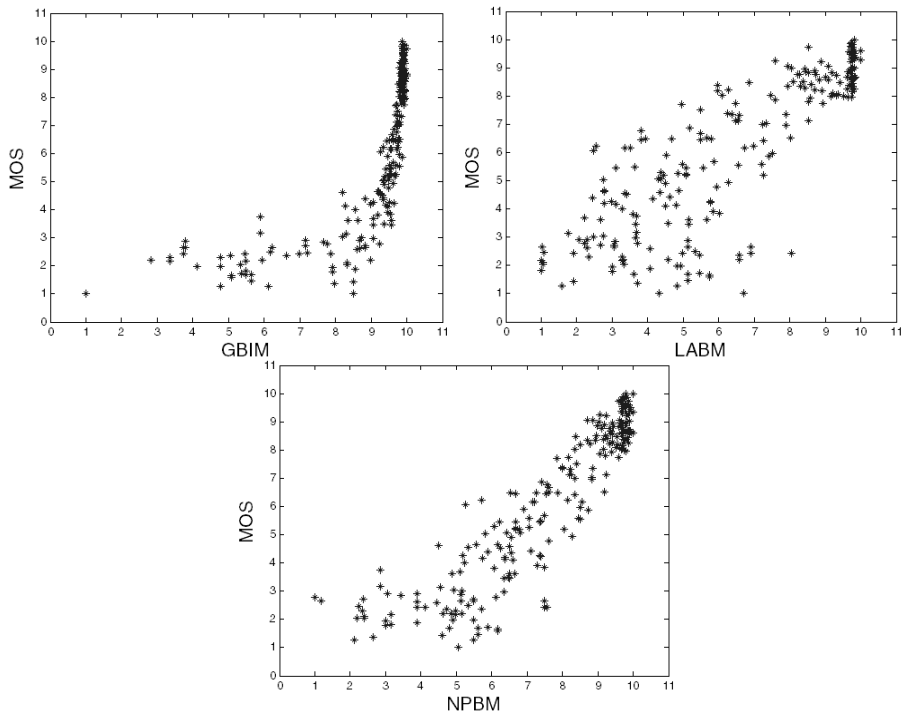


Fig. 10. Scatter plots of MOS vs. blockiness metrics.

Metric	Pearson Linear Correlation	Spearman Rank Order Correlation	Outlier Ratio
GBIM	0.790	0.912	0.099
LABM	0.834	0.832	0.009
NPBM	0.918	0.924	0

Table. 1. Performance comparison of three blockiness metrics.

The proposed overall blockiness metric, NPBM, is compared to state-of-the-art no-reference blockiness metrics based on a HVS model, namely GBIM [3] and LABM [4]. All three metrics are applied to the LIVE database of 233 JPEG images, and their performance is characterized by the linear correlation coefficients between the subjective MOS scores and the objective metric results. Figure 10 shows the scatter plots of the MOS versus GBIM, LABM and NPBM, respectively. The corresponding correlation results are listed in Table 1. It should be emphasized again that the correlation coefficients would be higher when allowing for a nonlinear mapping of the results of the metric to the subjective MOS. To illustrate the effect, the correlation coefficients were recalculated after applying the non-linear mapping function recommended by VQEG [21]. In this case, GBIM, LABM,

and NPBM yield a Pearson correlation coefficient of **0.928**, **0.933** and **0.946**, respectively.

GBIM manifests the lowest prediction accuracy among these metrics. This is mainly due to its human vision model used, which has difficulties in handling images under demanding circumstances, e.g. the highly textured images in the LIVE database. LABM adopts a more flexible HVS model, i.e. the JND profile with a more efficient integration of luminance and texture masking. As a consequence, the estimation of artifact visibility is more accurate for LABM than for GBIM. Additionally, LABM is based on a local estimation of blockiness, in which the distortion and its visibility due to masking are measured for each individual coding block of an image. This locally adaptive algorithm is potentially more accurate in the production of an overall blockiness score. In comparison with GBIM and LABM, our metric NPBM shows the highest prediction ability. This is primarily achieved by the combination of a refined local metric and a more efficient model of visual masking, both considering the specific structure of the artifact itself.

2.4 Evaluation of Specific Metric Components

The blocking annoyance metric, proposed in this paper, is primarily based on three aspects: (1) a grid detector to ensure the subsequent *local* processing; (2) a *local* distortion measure; and (3) a HVS model for *local visibility*. To validate the added value of these aspects, additional experiments were conducted and a comprehensive comparison to alternatives is reported. This includes a comparison of:

- metrics with and without a grid detector
- the local versus global approach
- metrics with and without a HVS model
- different HVS models

2.4.1 Metrics with and without a Grid Detector

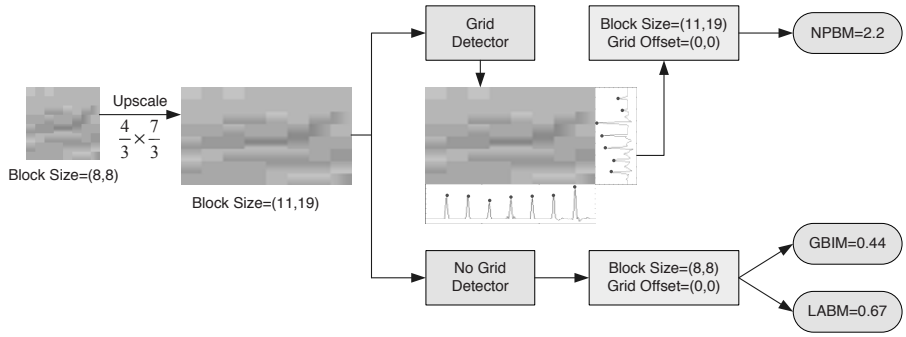


Fig. 11. Illustration of how to evaluate the effect of a grid detector on a blockiness metric: an image patch showing visible blocking artifacts was up-scaled with a scaling factor 4/ 3 x 7/ 3, and the metrics NPBM, GBIM and LABM were applied to assess the blocking annoyance of the scaled image.

Our metric includes a grid detection algorithm to determine the exact location of the blocking artifacts, and thus to ensure the calculation of the metric at the appropriate pixel positions. It avoids the risk of estimating blockiness at wrong pixel positions, e.g. in scaled images. To illustrate the problem of blockiness estimation in scaled images a small experiment was conducted. As illustrated in Figure 11, an image patch of 64x64 pixels was extracted from a low bit rate (0.34bpp) JPEG image of the LIVE database. This image patch had a grid of blocks of 8x8 pixels starting at its top-left corner, and it clearly exhibited visible blocking artifacts. It was scaled up with a factor 4/ 3 x 7/ 3, resulting in an image with an effective block size of 11x19 pixels. Blocking annoyance in this scaled image was estimated with three metrics, i.e. NPBM, GBIM and LABM. Due to the presence of a grid detector, the NPBM yielded a reasonable score of **2.2** (NPBM scores range from 0 (no blockiness) to 10 for the highest blocking annoyance). However, in the absence of a grid detector, both GBIM and LABM didn't detect any substantial blockiness: they had a score of GBIM=**0.44** and LABM=**0.67**, which corresponds to "no blockiness" according to their scoring scale (see [3] and [4]). Thus, GBIM and LABM fail in predicting blocking annoyance of scaled images, mainly due to the absence of a grid detector. Clearly these metrics could benefit in a similar way as our own metric from including the location of the grid.

Various alternative grid detectors are available in literature. They all rely on the gradient image to detect the blocking grid. To do so, they either calculate the FFT for each single row and column of an image [8], or they calculate the normalized gradient for every pixel in its two dimensions [25]. Especially, for large images (e.g. in the case of HD-TV), these operations are computationally expensive. The main advantage of our proposed grid detector lies in its simplicity, compared to existing alternatives in literature. Such as in the approach reported in [11], we first project the gradient image into a 1-D signal, and then enhance the signal maxima using once a median filter. In addition, the size and offset of the grid are extracted from the resulting 1-D signal using a DFT. The latter is less computationally expensive than the approach chosen in [11], being a complex maximum-likelihood method.

Apart from affecting the blocking grid position, scaling may also affect the blocking artifact visibility [25]. This aspect, however, is not yet taken into account in our proposed metric.

2.4.2 Local versus Global Approach

The difference in local versus global approach can be best understood by comparing their basic formulation. A local metric, as proposed in this paper, is based on a general formulation of the form MF1:

$$MF1 = \frac{1}{n} \sum_{k=1}^{n} [\, LPB(k) \times M(k) \,] \tag{19}$$

where k denotes the pixel location of blocking artifacts, and LPB and M denote the local pixel-based blockiness (see equation (8)) and the HVS model embedded, respectively. Both of them are calculated locally within a region of the image centered on individual blocking artifacts.

A global metric, as e.g. used in GBIM [3] is based on a general formulation of the form MF2:

$$MF2 = \frac{\left\| G(i,j)_{block-edge} \times M(i,j)_{block-edge} \right\|}{\left\| G(i,j)_{non-block-edge} \times M(i,j)_{non-block-edge} \right\|} \tag{20}$$

where G denotes the inter-pixel difference (see equation (1)), M denotes the HVS model embedded, and $||\ .\ ||$ is the L2-norm. The numerator is calculated at the location of blocking artifacts, while the denominator is calculated for pixels which are not on the blocking grid.

An obvious advantage of the local approach over the global approach is already revealed by their formulation: MF1 only calculates the HVS model for pixels on the blocking grid, while MF2 needs to calculate the HVS model for all pixels in the image. Since the major cost of a HVS-based blockiness metric is usually introduced by the human vision model, reducing the number of times the HVS model is calculated in the whole process is highly beneficial for the computational load. The computational cost related to the number of times the HVS model has to be calculated in a metric can be quantified by means of a *model utilization ratio* (MUR) , which is simply defined as the total number of times T_M that the HVS model is computed, divided over the total number of pixels MxN in the image

$$MUR = \frac{T_M}{M \times N} \tag{21}$$

Evidently, the lower this ratio, the simpler the metric is.

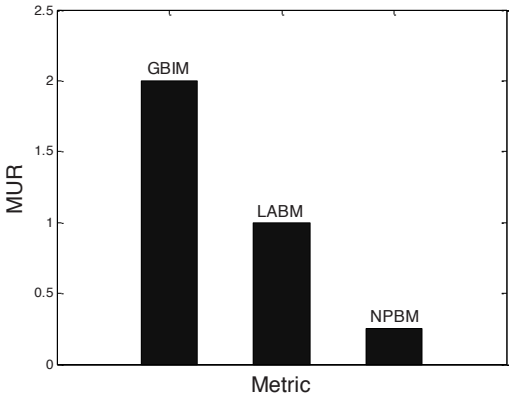


Fig. 12. Comparison of the computational cost of three metrics, using model utilization ratio (MUR).

Figure 12 shows the MUR for GBIM, LABM, and NPBM, respectively. Both GBIM and LABM calculate the human vision model for every pixel in an image, which yields a MUR of **1**. For GBIM the MUR is increased by a factor of **2**, since masking is estimated for the horizontal and vertical blockiness direction separately.

For our metric the MUR is only **0.25** in case of a block size of 8x8 pixels, which is a direct result of calculating the HVS model only at detected blocking artifacts. This implies that when neglecting the difference in computational cost between the various HVS models for a moment, the computational load of NPBM is reduced by approximately **7/8** with respect to GBIM, and by **3/4** with respect to LABM.

Of course, in this respect also the complexity of the HVS model used needs to be taken into account. This is further discussed in Section IV.D, taking into account various HVS models. Additionally, there also is a performance difference between the local and global approach. But, since the performance gain depends on the specific choice of HVS used, this point is also discussed in Section IV.D.

2.4.3 Metrics with and without a HVS model

To validate the added value of including a HVS model in a blockiness metric, we compared our proposed HVS-based metric NPBM to the state-of-the-art non-HVS-based metric of [25], which is referred to as NBAM. NBAM is also a global metric formulated according to equation (20), but instead of using a HVS model, it replaces the inter-pixel difference by the relative gradient in order to determine the visual strength of a block discontinuity. It was achieved a promising performance over the entire LIVE database as indicated by the Pearson correlation coefficient (after nonlinear regression) of **0.92**, which is comparable to our metric with a Pearson correlation coefficient of **0.94**. However, because of the absence of a HVS model, the robustness of NBAM against image content might be an issue. It may be doubted to what extent the objective metric is able to predict blockiness in more demanding images, e.g., for a set of highly textured images, compressed at very low bit-rates, for which visual masking is important.

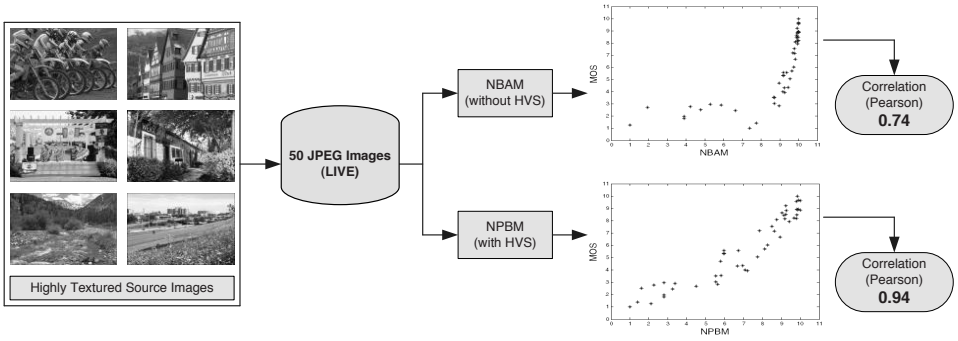


Fig. 13. Illustration of the added value of including a HVS model in a blockiness metric: a database of 50 highly textured JPEG images was extracted from the LIVE database, and blockiness annoyance was estimated with the metrics NBAM (without HVS) and NPBM (with HVS). The prediction performance is given in terms of the Pearson correlation coefficient.

To evaluate this, a subset of six highly-textured images, as shown in Figure 13, was selected from the twenty-nice source images of the LIVE database. Including different compression levels, this resulted in a test database of 50 JPEG images with

their corresponding MOS score extracted from the LIVE database. For these images, texture masking was dominant, i.e., most blocking artifacts were largely masked by background non-uniformity.

The blockiness metrics, NPBM and NBAM, were applied to this test database. Their prediction performance is quantified by the Pearson correlation coefficient (without nonlinear regression) as illustrated in Figure 13. As expected, the simple metric NBAM fails in accurately predicting the subjective ratings of this subset of demanding images, mainly due to the lack of a HVS model. NPBM shows a robust prediction ability, resulting in a high correlation with the subjective MOS.

2.4.4 Comparison of Different HVS Models

To compare the added value of our proposed HVS model to existing alternatives, various HVS models M have been embedded in the general formulation of our local metric (see MF1 in equation (19)). For M we used four alternatives:

- VC model (i.e. our proposed HVS model);
- JND model (i.e. the JND profile model based on [18]);
- WF model (i.e. the HVS model used in GBIM [3]);
- M=1 model (i.e. no HVS model embedded).

Doing so, resulted in four blockiness metrics, which we refer to as LM_{VC} (i.e. NPBM), LM_{JND}, LM_{WF} and LM_{NO}, respectively. These four metrics were applied to the LIVE database of 233 JPEG images. The metric performance was quantified by the Pearson correlation coefficient (without nonlinear regression) as illustrated in Figure 14. In such a scenario, the performance difference between any two metrics can be attributed to the HVS model embedded. LM_{NO} (i.e. MF1 without any HVS model) is used as the benchmark, and the HVS model gain is determined by calculating the difference in Pearson correlation coefficient between the metric LM_{NO} and any of the other three metrics.

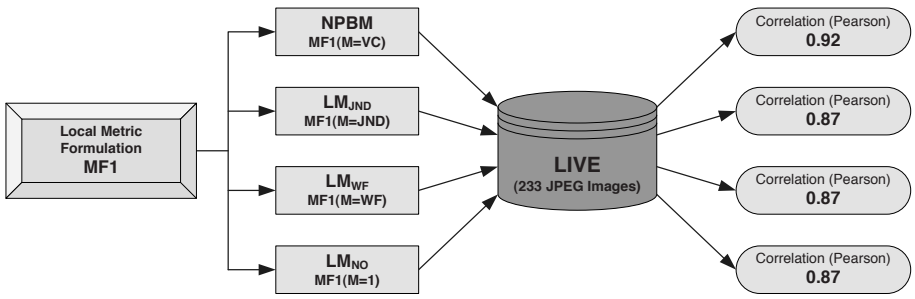


Fig. 14. Illustration of the comparison of various HVS models: a blockiness metric (i.e. MF1) having four optional HVS models embedded is tested with the LIVE database, and the performance for each resulting metric is quantified by the Pearson correlation coefficient.

Figure 14 clearly illustrates that our HVS model yields the biggest gain compared to the other three alternatives. For the local approach defined as MF1 in equation (19), there is no added value of using the JND or WF model in the metric, since their performance is comparable to that of the metric without HVS model. This may, of course, be due to the fact that the JND and WF model were not designed to be

combined with our proposed local metric. Our VC model, on the other hand, is designed together with the definition of MF1, and as a result a high correlation coefficient is found for the NPBM metric.

To investigate whether our HVS model is also valuable for traditionally used global metrics (see MF2 in equation (20)), the same experiment was repeated by substituting in MF2 the four options for M. This yielded another set of four blockiness metrics, which are referred to as GM_{VC}, GM_{JND}, GM_{WF} (i.e. GBIM), and GM_{NO}, respectively. Their performance when applied to the LIVE database is illustrated in Figure 15.

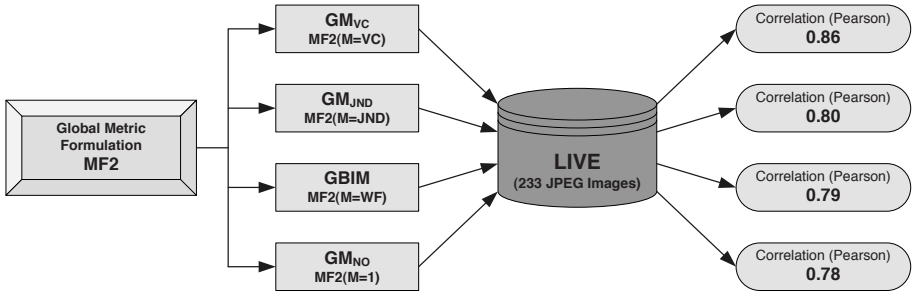


Fig. 15. Illustration of the comparison of various HVS models: a blockiness metric (i.e. MF2) having four optional HVS models embedded is tested with the LIVE database, and the performance for each resulting metric is quantified by the Pearson correlation coefficient.

It illustrates that also for a global metric our HVS model has the largest added value. In this case, however, also the WF and JND model have some added value. It should be noted, however, that in our evaluations the WF and JND model were implemented as described in the original publications (i.e. [3] and [18]). Some parameters in the implementations may be adjusted specifically to the LIVE database to provide a better correlation.

To summarize, the contribution of our proposed HVS model to a blockiness metric is consistently shown, independent of the specific design of the blockiness metric. In addition, a number of significant simplifications used in our HVS model are already discussed in Section II.C. The complexity of our VC model is comparable to that of the WF model, both of them use a simple weighting function for local visibility. However, the JND model is a rather complex HVS model, mainly due to the difficulties in estimating the visibility thresholds for various masking effects, and in combing different JND thresholds. The simplicity of the VC model itself, coupled with its specific design for a local approach to avoid calculating it on irrelevant pixels, consequently make this HVS model especially promising in terms of real time applications.

An additional interesting finding from the comparison of Figures 14 and 15 is that there is indeed a gain in performance applying the MF1 formulation (local approach) instead of the MF2 formulation (global approach), independent of the HVS model used. In the absence of any HVS model, the gain of MF1 over MF2 (i.e. from LM_{NO} to GM_{NO}) corresponds to an increase in the Pearson correlation coefficient from **0.78** to **0.87**. For the other HVS models, the corresponding numbers

are summarized in Figure 16. It confirms that a promising performance is achieved when applying the *local* approach in a blockiness metric.

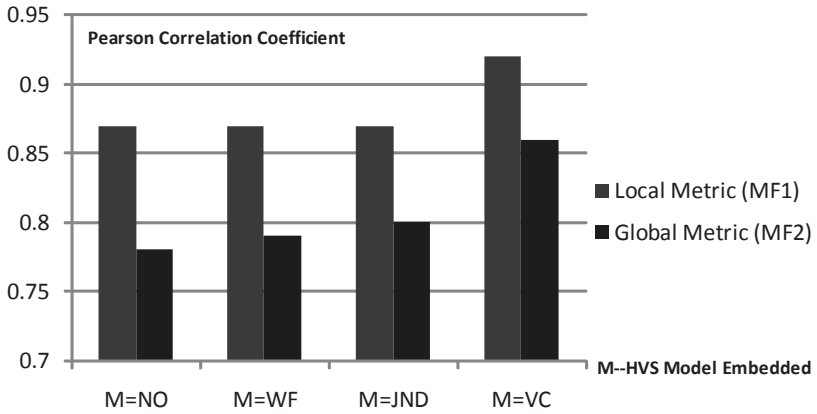


Fig. 16. Comparison of the local and global approaches to a blockiness metric, and of metrics with different HVS models embedded.

2.5 Conclusions

In this paper, a novel blockiness metric to assess blocking annoyance in block-based DCT coding is proposed. It is based on the following features:
- a simple grid detector to ensure the effectiveness of the blockiness metric, and to account for deviations in the blocking grid of the incoming signal or as a consequence of spatial scaling.
- a local pixel-based blockiness value that measures the strength of the distortion within a region of the image centered around each individual blocking artifact.
- a simplified and more efficient model of visual masking, exhibiting an improved robustness in terms of content independency, and allowing supra-threshold estimation of perceived annoyance.

An advantage of the proposed approach, especially in case of real-time application, is that the additional computational cost introduced by the HVS is largely reduced by eliminating calculations of the human vision model for non-relevant pixels. This is primarily accomplished taking advantage of the locality of both the pixel-based blockiness value and the visibility model. Nonetheless, the metric is mainly used to assess overall blockiness annoyance, which is simply done by summing the local contributions over the whole image.

Experimental results show that our proposed blockiness metric results in a strong correlation with subjective data, and outperforms state-of-the-art metrics in terms of prediction accuracy. Combined with its practical reliability and computational efficiency, our metric is a good alternative for real-time implementation.

2.6 References

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

[2] C. C. Koh, S. K. Mitra, J. M. Foley, and I. Heynderickx, "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, March 2005.

[3] H. R. Wu and M. Yuen, "A Generalized Block-edge Impairment Metric for Video Coding," IEEE Signal Processing Letters, vol. 70, no. 3, pp. 247-278, Nov. 1998.

[4] F. Pan, X. Lin, S. Rahardja, W. Lin, E. Ong, S. Yao, Z. Lu, and X. Yang, "A locally adaptive algorithm for measuring blocking artifacts in images and videos," Signal Processing: Image Communication 19 (6) (2004) 499-506.

[5] S. Winkler, "Issues in Vision Modeling for Perceptual Video Quality Assessment," Signal Processing, vol. 78, no. 2, pp. 231-252, Oct. 1999.

[6] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in Proc. IEEE Int. Conf. Image Processing, vol. 1, pp. 477-480, Sept. 2002.

[7] R. V. Babu, S. Suresh, and A. Perkis, "No-reference JPEG-image quality assessment using GAP-RBF," Signal Processing 87 (6): 1493-1503, 2007.

[8] Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in image," in Proc. IEEE Int. Conf. Image Processing, vol. 3, pp. 981-984, Sep. 2000.

[9] S. Liu and A. C. Bovik, "Efficient DCT-Domain Blind Measurement and Reduction of Blocking Artifacts," IEEE Trans. on Circuits and Systems for Video Technology, vol. 12, no. 12, pp. 1139-1149, Dec. 2002.

[10] E. Lesellier and J. Jung, "Robust wavelet-based arbitrary grid detection for MPEG," in Proc. IEEE Int. Conf. Image Processing, vol. 3, pp 417-420, 2002.

[11] S. Tjoa, W. S. Lin, H. V. Zhao, and K. J. R. Liu, "Block Size Forensic Analysis in Digital Images," in Proc. IEEE Int. Conf. ICASSP, vol. 1, pp. 633-636, 2007.

[12] S. Winkler, "Vision Models and Quality Metrics for Image Processing Applications," Ph.D. dissertation, Dept. Elect., EPFL, Lausanne, 2002.

[13] Z. Yu and H. R. Wu, "Human Visual System Based Objective Digital Video Quality Metrics," In Proc. Int. Conf. Signal Processing, vol. II, pp.1088-1095, Aug. 2000.

[14] Z. Yu, H. R. Wu, S. Winkler, and T. Chen, "Vision Model Based Impairment Metric to Evaluate Blocking Artifacts in Digital Video," in Proc. of the IEEE, pp. 154-169, Jan. 2002.

[15] E. M. Yeh, A. C. Kokaram, and N. G. Kingsburg, "A Perceptual Distortion Measure for Edge-Like Artifacts in Image Sequences," Human Vision and Electronic Imaging SPIE, vol. III, pp. 160-172, 1998.

[16] S. A. Karunasekera and N. G. Kingsbury, "A Distortion Measure for Blocking Artifacts in Images Based on Human Visual Sensitivity," IEEE Trans. Image Processing, vol. 4, no. 6, pp. 713-724, June 1995.

[17] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, "Motion-Compensated Residue Preprocessing in Video Coding Based on Just-Noticeable-Distortion Profile,"

IEEE Trans. on Circuits and Systems for Video Technology, vol. 15, no. 6, pp. 742-751, 2005.
[18] C. H. Chou and Y. C. Li, “A Perceptually Tuned Subband Image Coder Based on the Measure of Just-Noticeable-Distortion profile,” IEEE Trans. on Circuits and Systems for Video Technology, no. 6, pp. 467-476, Dec. 1995.
[19] T. N. Pappas and R. J. Safranek, Perceptual criteria for image quality evaluation. In Handbook of Image and Video Processing, Academic Press, May 2000.
[20] K. I. Laws, “Texture Energy Measures,” In Proc. DARPA Image Understanding Workshop, pp. 47-51, 1979.
[21] 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
[22] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. (2008, Mar. 21): LIVE image quality assessment database Release 2. Available: http:/ / live.ece.utexas.edu/ research/ quality
[23] B. Girod, “The information theoretical significance of spatial and temporal masking in video signals,” in Proc. SPIE Conf. Human Vision, Visual Processing, and Digital Display, vol. 1077, pp. 178-187, 1989.
[24] M. Yuen and H. R. Wu, “A survey of hybrid MC/ DPCM/ DCT video coding distortions,” Signal Processing, 70(1998) 247-278, 1998.
[25] R. Muijs and I. Kirenko, “A No-Reference Blocking Artifact Measure for Adaptive Video Processing,” in Proc. 13th European Signal Processing Conference, Turkey, 2005.
[26] I. O. Kirenko, R. Muijs, and L. Shao, “Coding Artifact Reduction using Non-Reference Block Grid Visibility Measure,” in Proc. IEEE Int. Conf. Multimedia and Expo, pp. 469-472, 2006.
[27] H. Liu and I. Heynderickx, "A No-Reference Perceptual Blockiness Metric", in Proc. IEEE Int. Conf. ICASSP, pp. 865-868, March 2008.

Chapter 3

A Perceptually Relevant Approach to Ringing Region Detection

***Abstract*: An efficient approach towards a no-reference ringing metric intrinsically exists of two steps: first detecting regions in an image where ringing might occur, and second quantifying the ringing annoyance in these regions. This paper presents a novel approach towards the first step: the automatic detection of regions visually impaired by ringing artifacts in compressed images. It is a no-reference approach, taking into account the specific physical structure of ringing artifacts combined with properties of the human visual system (HVS). To maintain low complexity for real-time applications, the proposed approach adopts a perceptually relevant edge detector to capture regions in the image susceptible to ringing, and a simple yet efficient model of visual masking to determine ringing visibility. The approach is validated with the results of a psychovisual experiment, and its performance is compared to existing alternatives in literature for ringing region detection. Experimental results show that our method is promising in terms of both reliability and computational efficiency.**

This chapter is based on the research article published as "A Perceptually Relevant Approach to Ringing Region Detection" by H. Liu, N. Klomp and I. Heynderickx in IEEE Transactions on Image Processing, vol. 19, pp. 1414-1426, June, 2010.

3.1 Introduction

In current visual communication systems, the most essential task is to fit a large amount of visual information into the narrow bandwidth of transmission channels or into a limited storage space, while maintaining the best possible perceived quality for the viewer [1]. A variety of compression algorithms, such as e.g. JPEG and MPEG/ H.26x, have been widely adopted in image and video coding trying to achieve high compression efficiency at high quality [2], [3]. These lossy compression techniques, however, inevitably result in various coding artifacts, which by now are known and classified as blockiness, ringing, blur, etc. [4]. The occurrence of the compression induced artifacts depends on the data source, target bit-rate, and underlying compression scheme, and their visibility can range from imperceptible to very annoying, thus affecting perceived quality [5], [6], [7]. During the last decades a lot of research effort is devoted to reduce coding artifacts, so to improve the overall perceived quality of artifact impaired image material [8], [9], [10]. In the video chain of a current TV-set e.g., various video enhancement algorithms, such as de-blocking, de-ringing and de-blur, are typically employed to reduce compression artifacts prior to display. In such a scenario, objective metrics, which determine the quality degradation caused by each individual artifact, and adapt the processing chain for artifact reduction accordingly, are highly needed. In addition, the receiving end of a digital video chain usually has no access to the original image, and in most cases there is even only limited access to the encoding parameters of the bit-stream. Hence, objective metrics used in these types of applications are constrained to a no-reference approach, which means that the impairment assessment relies on the compressed image only.

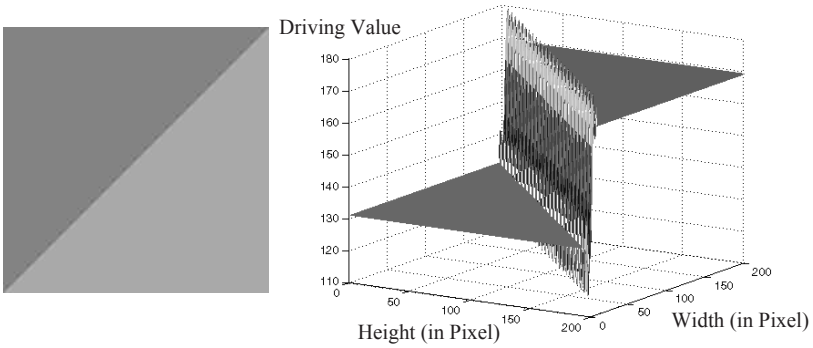


Fig. 1. Illustration of ringing artifacts in an image patch compressed with JPEG (MATLAB's *imwrite* function with Q=30): (a) 2D image and (b) its spatial intensity distribution (in 8-bits driving values). Ringing can be perceived as intensity fluctuations near the edges, while the image content there should be uniform.

In the last decades, a considerable amount of research has been devoted to the development of a blockiness metric (see e.g. [11] and [12]), which has been already implemented for the optimization of image quality (see e.g. [13], [14], [15]). Another common distortion type, namely ringing [4], intrinsically results from loss in the

high frequency component of the video signal due to coarse quantization. In the spatial domain, ringing, which is fundamentally associated with Gibb's phenomenon, manifests itself in the form of ripples or oscillations around high contrast edges. The occurrence of ringing artifacts spreads out to a finite extent surrounding edges, depending on the underlying properties of the compression scheme. For example, in block-based DCT coding ringing appears as a ripple outwards from the edge up to the encompassing block's boundary [4]. As an example, Figure 1 illustrates ringing artifacts induced by JPEG compression.

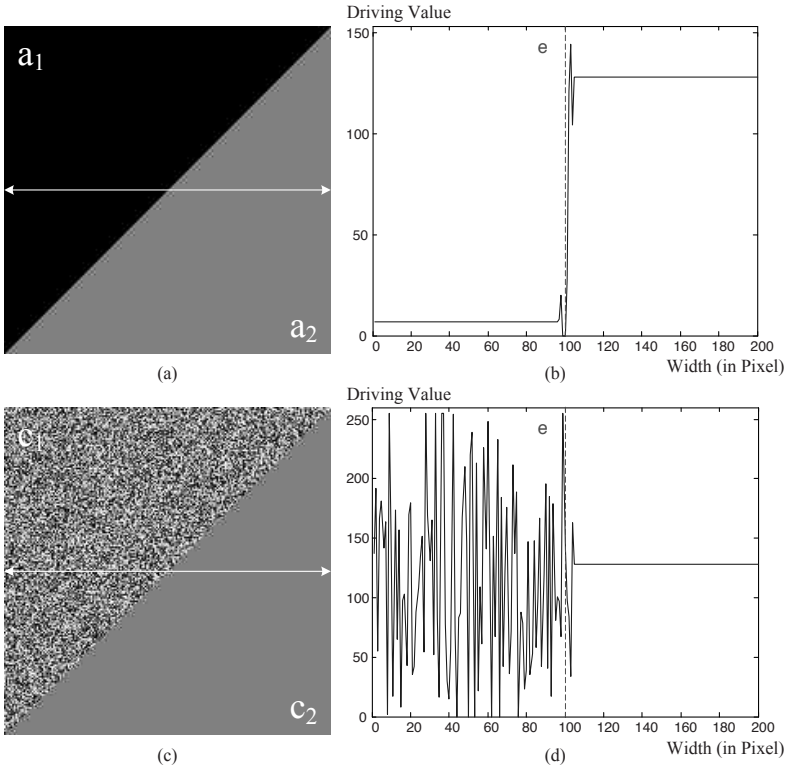


Fig. 2. Illustration of luminance ((a) and (b)) and texture ((c) and (d)) masking on ringing visibility for two image patches compressed with JPEG (MATLAB's *imwrite* function with Q=30). Graphs (a) and (c) show the compressed image patches, whereas graphs (b) and (d) represent the intensity profile (in 8-bits driving values) along the row in the image indicated with the arrow in graphs (a) and (c), respectively. The dashed line "e" in graphs (b) and (d) refers to the position of the edge. Note that although both sides of the edge at "e" exhibit ringing artifacts, the visibility of ringing differs.

Research on the design of a blockiness metric has shown that an efficient no-reference approach intrinsically exists of two steps: (1) the detection of regions in an image where blockiness might occur, and (2) the determination of the blocking annoyance in these regions. We use a similar two-step approach for the design of a no-reference ringing metric. This paper only discusses the first step: the detection of regions in the image, in which visible ringing occurs. A successive paper that

discusses the quantification of the perceived annoyance of ringing in these regions is published in [16] and [17].

Unlike blocking, whose spatial location is very regular and thus easily predictable, the location of ringing is edge dependent, and as such also image content dependent. This makes the task of detecting ringing regions much more difficult, especially in a no-reference application. In general, ringing can be considered as a form of signal dependent noise, which only occurs near sharp transitions in image intensity when not visually masked by local image characteristics. As such, the occurrence of ringing can be directly associated with strong edges in an image. Additionally, the visibility of ringing is reduced in the case of very low and very high background intensity (i.e. luminance masking [18]), and ringing is more visible in homogenous areas than in textured or detailed areas (i.e. texture masking [19]). The effect of luminance and texture masking on ringing visibility is illustrated in Figure 2. Hence, to accurately detect regions with perceived ringing, two essential aspects need to be explicitly addressed: (1) an (strong) edge detector; and (2) a masking model of the HVS.

3.1.1 Review of Related Work

Until recently, only a limited amount of research was devoted to perceived ringing. The methods in [20] and [21] both simply assume that ringing occurs unconditionally in regions surrounding strong edges in an image. This, however, does not always reflect human visual perception of ringing, because of the absence of spatial masking as typically present in the HVS. This issue is taken into account by incorporating properties of the HVS into the detection method, such as e.g. in [22] and [23]. The approach in [22] is based on the global edge map of an image, where binary morphological operators are used to generate a mask to expose regions that are likely to be contaminated with visible ringing artifacts. This procedure involves the identification of regions around all detected edges, and a further evaluation of these regions based on visual masking. In [23], a different way of including HVS masking properties is employed. This method classifies the potential smooth regions (i.e. regions in an image other than edges and their surroundings) into different objects based on their color similarity and texture features. The resulting objects are assigned as background around potential ringing regions. Texture masking is implemented by evaluating the contrast in activity between the potential ringing region and its assigned background (e.g. the higher the contrast in activity, the more visible ringing is assumed to be). Additionally, also luminance masking is implemented to further determine ringing visibility.

There are two main concerns with the methods existing in literature. First of all, the edge detection methods employed in [20], [21], [22], [23] capture strong edges using an ordinary edge detector, such as a Sobel operator, where a certain threshold is applied to the gradient magnitudes to remove noise and insignificant edges. Depending on the choice of the threshold, these methods run the risk of omitting obvious ringing regions near non-detected edges (in case of a high threshold) or of increasing the computational power by modeling the HVS near irrelevant edges (in case of a low threshold). Figure 3 illustrates the effect of the threshold value of a Sobel operator. The edge map in Figure 3(c), resulting from a high threshold value, largely removes noisy edges while eliminating a number of important edges, at

which ringing obviously exists (see Figure 3(b)). This may heavily degrade the accuracy of the prediction of perceived ringing. By lowering the threshold (as in Figure 3(d)), all strong edges are maintained in the edge map, but it also contains more texture edges, which are non-relevant to ringing detection, and consequently, result in a large number of unnecessary computations for ringing visibility. The second concern with the existing methods is related to the models of the HVS used e.g. in [22] and [23], which are computationally very expensive. The HVS model in [22] involves a parameter estimation mechanism, which requires a number of calculations to achieve an optimal selection. The major cost of the HVS model in [23] is introduced by its clustering scheme embedded, which contains color clustering and texture clustering.

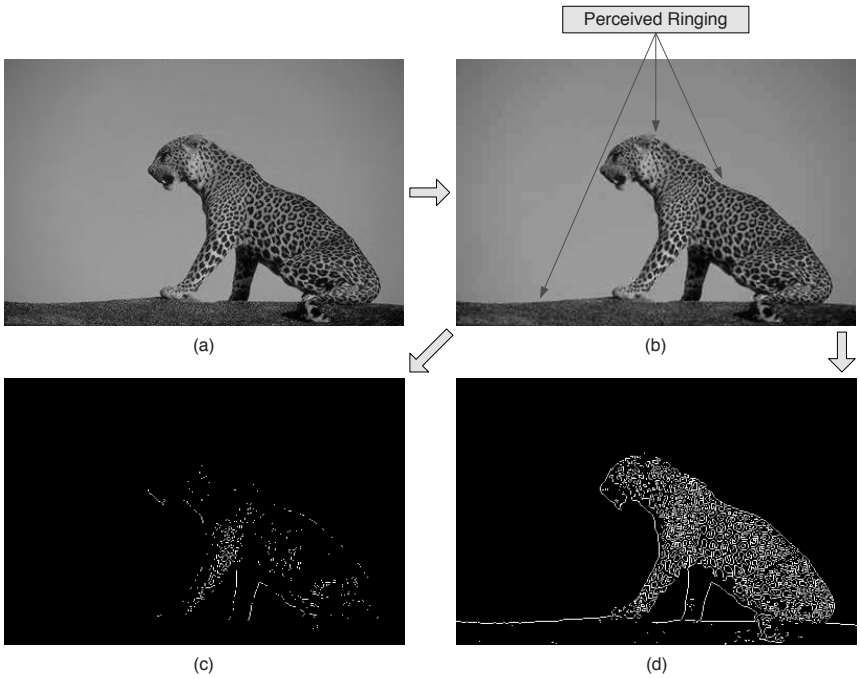


Fig. 3. An ordinary edge detector (i.e. Sobel operator) applied for ringing region detection: (a) original image, (b) JPEG compressed image (MATLAB's *imwrite* function with Q= 30), (c) Sobel edge map of (b) using a high threshold (i.e. 23% of I_{max}), and (d) Sobel edge map of (b) using a low threshold (i.e. 10% of I_{max}).

Obviously, the optimal performance in terms of reducing the number of required computations, while maintaining the reliable detection of perceived ringing, can be achieved by optimizing two aspects: (1) the detection accuracy of relevant edges; and (2) the reduction in complexity of the HVS model itself. Hence, what is needed is an edge detector that only extracts edges most closely related to the occurrence of ringing, and a HVS model that is simpler (and thus more applicable for real-time implementation) than the approaches existing in literature. In this paper, both aspects needed to efficiently detect regions with visible ringing are discussed.

3.2 Proposed Algorithm

The schematic overview of the proposed algorithm is illustrated in Figure 4. It mainly consists of two parts: (1) extraction of edges relevant for ringing, and (2) detection of visibility of ringing in the edge regions. In the first part, an advanced edge detector is adopted, attempting to select the edges most relevant for ringing (i.e. contours of objects) in combination with the avoidance of the irrelevant edges (i.e. in textured areas). This results in a perceptual edge map (PEM), existing of a set of so-called line segments (LS). In the second part, each LS of the PEM is examined individually on the occurrence of visible ringing in its direct neighborhood, taking into account masking by the HVS. All regions with visible ringing are accumulated in a single binary map, which we refer to as the computational ringing region (CRR) map. Remind that the CRR map is used as input to the second step of the objective metric, in which the ringing annoyance is *quantified*, as published in [16] and [17]. Each part of the ringing region detection algorithm is further detailed in the following sections. The parameters used in the algorithm are specified and discussed in Section IV.B. Note that the entire metric is only based on the luminance channel of the images in order to further reduce the computational load.

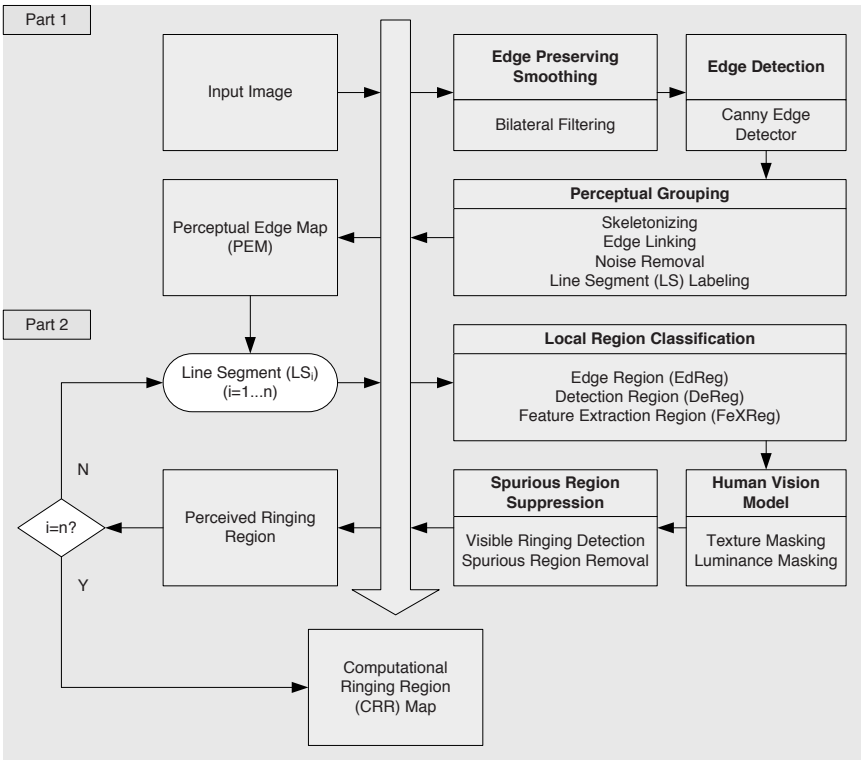


Fig. 4. Schematic overview of the proposed algorithm, with at the top the part to detect edges relevant for ringing, and at the bottom the part to measure visibility of ringing around these edges.

3.2.1 Perceptual Edge Extraction

As explained above, the detection of visible ringing heavily relies on the accurate and efficient detection of object edges. To achieve this, we propose the application of a Canny edge detector [24] to an image, which first is non-linearly smoothened. After some additional post-processing, this results in the PEM.

Edge Preserving Smoothing and Canny Edge Detection

When interpreting the surrounding world, humans tend to respond to differences between homogeneous regions rather than to structure within these homogeneous regions [25]. Hence, finding perceptually strong edges mainly implies that texture existing in homogenous regions can be neglected as if viewed from a long distance. This can be implemented by smoothing the image progressively until textual details are significantly reduced, and then applying an edge detector.

Traditional low-pass linear filtering (e.g. Gaussian filtering) smoothens out noise and texture, but also blurs edges, and consequently, changes their spatial location. Since ringing detection intrinsically requires accurate spatial localization of the edges, edge-preserving smoothing is needed. Bilateral filtering was introduced in [26] as a simple and fast scheme for edge-preserving smoothing. It is a nonlinear operation that combines nearby image values based on both their geometric closeness and their photometric similarity, and prefers near values to distant values in both spatial domain and intensity range. In the Gaussian case, it can be expressed as:

$$\vec{F}(\vec{x}) = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \vec{I}(\vec{\xi})\omega(\vec{\xi},\vec{x})d\vec{\xi}}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \omega(\vec{\xi},\vec{x})d\vec{\xi}} \tag{1}$$

where

$$\omega(\vec{\xi},\vec{x}) = \exp\left(\frac{-(\vec{\xi}-\vec{x})^2}{2\sigma_d^2}\right)\exp\left(\frac{-(I(\vec{\xi})-I(\vec{x}))^2}{2\sigma_r^2}\right) \tag{2}$$

I and F denote the input and output images, χ and ξ are space variables, and the standard deviations σ_d and σ_r characterize the domain and range filtering, respectively. The advantage of using bilateral filtering instead of Gaussian filtering for the localization specific detection of perceptually strong edges is illustrated in Figure 5.

Subsequently, a Canny edge detector is applied to the bilaterally filtered image to obtain the perceptually more meaningful edges. Since the input image is already filtered, the subsequent Canny algorithm is implemented without its inherent smoothing step, while keeping the other processing steps unchanged. The Canny edge detector uses two thresholds to detect strong and weak edges, and includes the weak edges in the output only if they are connected to strong edges. Their values is automatically set, depending on the image content.

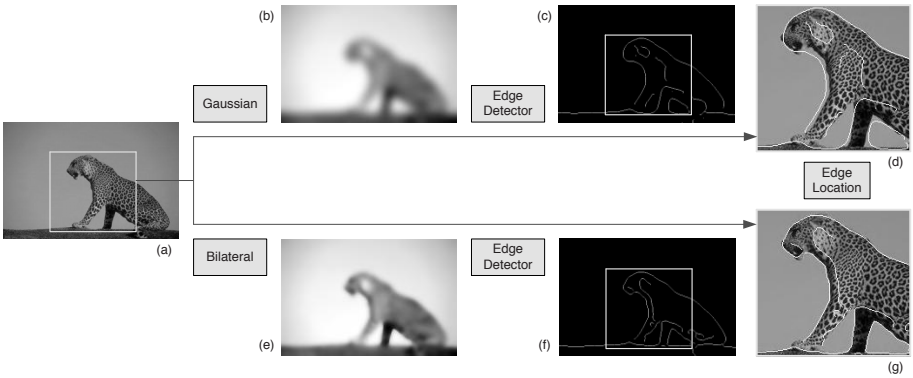


Fig. 5. Bilateral filtering and Gaussian filtering for the detection of perceptually strong edges: (a) original image, (b) Gaussian filtered image (σ_d=15), (c) edge map of (b), (d) superposition of (c) on (a), (e) bilateral filtered image (σ_d =3, σ_r =100), (f) edge map of (e), and (g) superposition of (f) on (a).

Perceptual Edge Map Formation

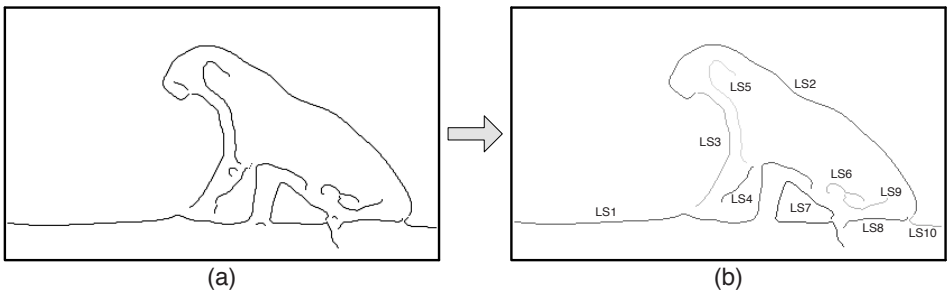


Fig. 6. Construction of the perceptual edge map (PEM): (a) Canny edge map and (b) related PEM with labeled line segments.

Since the HVS does not perceive luminance variations at pixel level, the detected edge pixels are necessarily combined into perceptually salient elements, facilitating further analysis and processing [25]. These perceptual elements, which we refer to as line segments (LS), are constructed over the Canny edge map and will be used as the basis for ringing region detection. The following processing steps are implemented to define the LS in the PEM.

1) Skeletonizing: To guarantee that an edge is only one-pixel thick, a kernel of 4x4 pixels is slid over all pixels, and those pixel configurations that have a structure of [1 1; 0 1] or [1 0; 1 1] are replaced by [1 0; 0 1], and those with a structure of [1 1; 1 0] or [0 1; 1 1] are replaced by [0 1; 1 0].

2) Edge Linking: The algorithm links all the edge pixels into a set of elements; each element either contains two end-points or is a closed loop. If an edge junction

is encountered, the tracing procedure breaks, and a separate element is generated for each of the branches.

3) Noise Removal: The elements with the number of connected edge pixels below a certain threshold are discarded. This is done with the ringing detection accuracy and speed in mind.

4) Line Segment Labeling: the resulting elements of connected edge pixels are referred to as line segments (LS), and labeled.

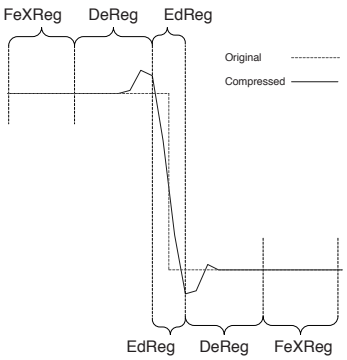
Once this process is complete, we have the PEM. Figure 6 illustrates the labeling of the LS in the PEM.

3.2.2 Ringing Region Detection

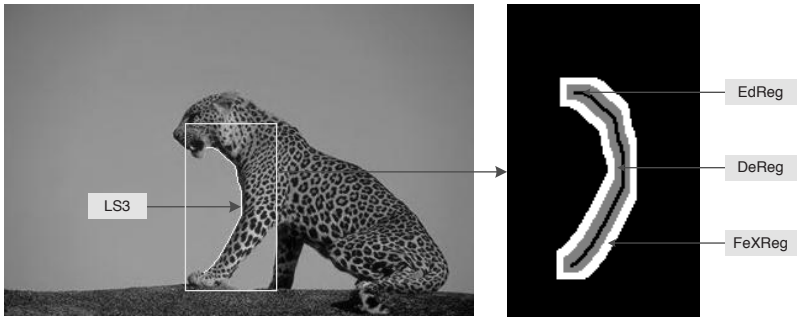
Each LS of the PEM is examined individually on the occurrence of visible ringing artifacts in their direct neighborhood, taking into account luminance and texture masking. The regions with visible ringing are then combined in a computational ringing region (CRR) map.

Local Region Classification

In order to characterize the visibility of ringing around a LS, its surrounding is classified into three different zones (see Figure 7(a) for an example of a single step edge): (1) Edge Region (EdReg): the original edge including the compression induced blur; (2) Detection Region (DeReg): the direct neighborhood of the EdReg, which potentially contains ringing artifacts; and (3) Feature Extraction Region (FeXReg): a region representative for the original local background, which is located outwards from the corresponding DeReg. These regions are defined by thickening the LS with a different size for the structuring element of a dilation operation. Figure 7(b) gives an example, in which for one LS (i.e. LS3 of Figure 6(b)) the EdReg, DeReg, and FeXReg obtained with a square structuring element of 2, 9 and 17 pixels width, respectively, is shown.



(a)



(b)

Fig. 7. Illustration of local region classification: (a) illustration of the three zones for a schematic step edge, and (b) illustration of how the zones are defined around an actual line segment as part of a natural image. In (b) the black line indicates the EdReg, the gray area defines the DeReg, and the white area refers to the FeXReg.

The Human Vision Model

Whether ringing is actually visible in the DeReg strongly depends (because of masking in the HVS) on the content of the original background, here represented by the FeXReg. Hence, the visibility of ringing is evaluated for each LS by applying a model for texture and luminance masking, using the texture and luminance characteristics of the FeXReg. As a result, DeReg regions, in which ringing is visually masked are eliminated, and only the perceptually prominent DeReg ringing regions remain.

1) Texture Masking

The visibility of ringing is significantly affected by the spatial activity in its local background, i.e. ringing is visually masked when located in a textured region, while it is perceptually prominent against a smooth background [22], [23], [27] as illustrated in Figure 2. In this paper, texture masking is modeled classifying the FeXReg of each LS into "smooth" and "textured" objects, depending on the local background characteristics. The DeReg is segmented accordingly, and those DeReg regions of which the corresponding FeXReg is clustered as "textured" are removed. This approach intrinsically avoids explicit modeling of the HVS, and formulates texture masking as a simple yet efficient local pixel clustering procedure. The proposed scheme to implement this is illustrated in Figure 8(b). It generally involves the following steps:

(1) Calculating the local activity of the image content covered by the FeXReg by applying a global threshold to the gradient in pixel intensity to create a local binary map (LBM) of the FeXReg. This yields a profile of local pixel activities, and is formulated as

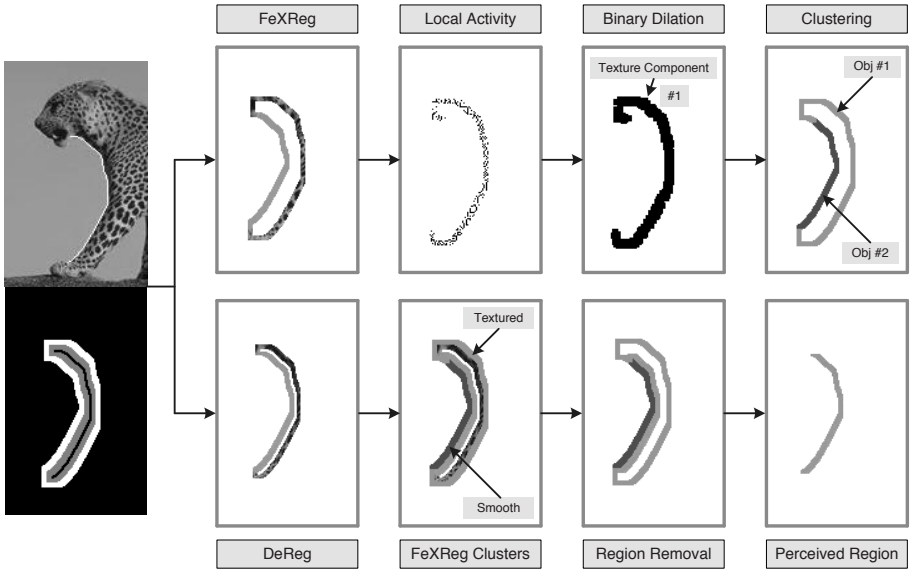
$$LBM(i,j)=\begin{cases}0 & LA(i,j)<Thr\_txt \\ 1 & otherwise\end{cases}, \qquad i,j\in \text{FeXReg} \tag{3}$$

$$LA(i,j)=\begin{vmatrix}[I(i-1,j-1)+2\times I(i-1,j)+I(i-1,j+1)]- \\ [I(i+1,j-1)+2\times I(i+1,j)+I(i+1,j+1)]\end{vmatrix}+ \\ \begin{vmatrix}[I(i-1,j+1)+2\times I(i,j+1)+I(i+1,j+1)]- \\ [I(i-1,j-1)+2\times I(i,j-1)+I(i+1,j-1)]\end{vmatrix} \tag{4}$$

where the local activity $LA(i, j)$ at location (i, j) is approximated by the gradient of the image intensity using a gradient operator (e.g. a Sobel operator). The 3x3 pseudo-convolution template used to calculate the gradient magnitude of a pixel at location (i, j) is shown in Figure 8(a) (I(i, j) corresponds to the pixel intensity at location (i, j)). The threshold *Thr_txt* is related to the magnitude histogram of the gradient image, and thus, image content dependent.

I (i-1, j-1)	I (i-1, j)	I (i-1, j+1)
I (i, j-1)	I (i, j)	I (i, j+1)
I (i+1, j-1)	I (i+1, j)	I (i+1, j+1)

(a)



(b)

Fig. 8. Implementation of texture masking: (a) Pseudo-convolution template used to calculate approximate gradient magnitude and (b) illustration of the algorithm.

(2) Dilating the LBM using a morphological operator, and labeling (e.g. by 8-connectivity) them into a set of connected components, which are referred to as texture components. This step intrinsically transfers pixel activities to a higher level structure of region activities, motivated by the fact that the human eye is not sensitive to variations at pixel level.

(3) Classifying all FeXReg covered by texture components into "texture objects", and the remaining FeXReg into "smooth objects".

(4) Removing the regions of DeReg that belong to the "texture objects" of FeXReg, since in these regions ringing is supposed to be masked by texture, and discarding the resulting regions of DeReg with their size under a certain threshold. The maintained regions of DeReg are considered as perceived ringing regions.

2) Luminance Masking

The visibility of variations in luminance depends on the local mean luminance [18], [19], [27], [28], [29]. As a result, the visibility of ringing is largely reduced in extremely dark or bright surroundings, as illustrated in Figure 2. The implementation of luminance masking is the same as for texture masking, but to guarantee efficiency, it is only applied to those regions of the DeReg remaining after the application of texture masking. The procedure for luminance masking is similarly formulated as a local pixel clustering model, and it mainly contains the following steps:

(1) Calculating the local averaged luminance, over a 3x3 template, centered on each pixel that is part of a "smooth object" of the FeXReg

$$LML(i,j) = \frac{1}{9}\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1} I(k,l) \tag{5}$$

where I(i, j) denotes the pixel intensity at location (i, j), and LML(i, j) denotes the local mean luminance. The visibility of ringing due to luminance masking is determined according to the functional behavior shown in Figure 9 [12], and a local binary map (LBM) is generated by applying a pre-defined threshold to the visibility coefficient (VC)

$$LBM(i,j) = \begin{cases} 0 & VC(i,j) > Thr\_lum \\ 1 & otherwise \end{cases} \tag{6}$$

where LBM(i, j)=0 indicates a visible pixel location, and LBM(i, j)=1 indicates a non-visible pixel location. This generates a profile of local visibility due to luminance masking.

(2) Dilating the LBM to obtain a set of connected components, which are referred to as invisible components.

(3) Classifying the "smooth objects" of FeXReg further into "visible objects" and "invisible objects" depending on the invisible components. This step combined with the one mentioned above intrinsically yields the structures of region visibility.

(4) Removing the DeReg that correspond to “invisible objects”, i.e. where ringing is not supposed to be visible against a very low or very high intensity background. Ultimately, only the regions of DeReg that yield visible ringing remain. These regions are combined in the CRR map, of which an example is given in Figure 10.

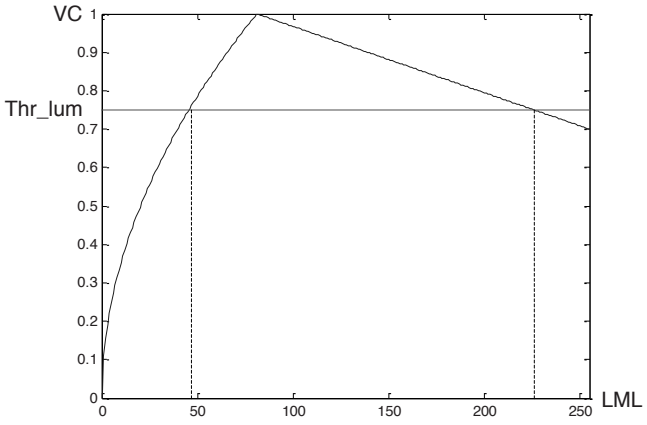


Fig. 9. Implementation of luminance masking via the relation between the local mean luminance (LML) and the artifact visibility coefficient (VC); Thr_lum refers to the threshold used in the implementation.

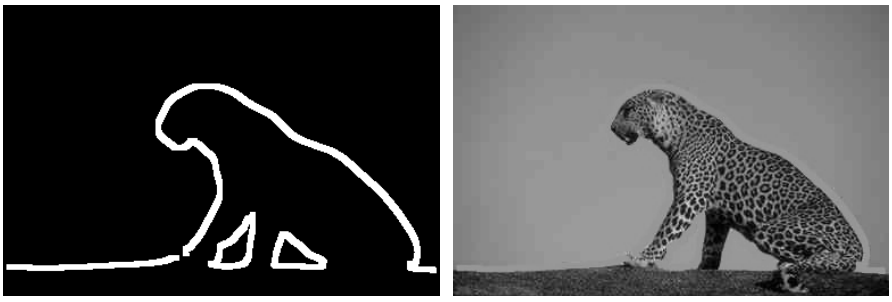


Fig. 10. Example of a computational ringing region (CRR) map (a) corresponding to a JPEG compressed image (b).

3.2.3 Spurious Ringing Region Suppression

The ringing region detection method described so far only exposes regions in an image which are likely to be impaired by visible ringing artifacts. The resulting CRR map, however, still includes obvious spurious ringing regions, containing either “unimpaired” or “noisy” pixels misinterpreted as ringing pixels.

“Unimpaired pixels” indicate pixels in the detected regions of the CRR map, which are actually not impaired by ringing. An obvious example of the occurrence of “unimpaired” pixels is in an uncompressed image. The ringing region detection

algorithm described so far will find the regions that might be impaired with visible ringing, independent of the compression level. But in an uncompressed image, these regions do not contain visible ringing, and hence, should be removed from the CRR map. Note that without removal of these regions the overall objective ringing metric including the step of quantification of ringing annoyance (see [16] and [17]) would not be less accurate, but less efficient.

"Noisy pixels" are pixels in the detected regions of the CRR map, that actually belong to an edge or texture. They are accidentally misclassified to a ringing region as a consequence of the dilation operation used in the human vision model.

To remove the spurious ringing regions, each detected ringing region (RR) is further examined by calculating its amount of visible ringing pixels. Those RRs with their number of visible ringing pixels below a certain threshold are considered as spurious, and consequently removed from the CRR map. Whether a pixel in a RR is a visible ringing pixel is determined via the local variance (LV) in intensity in its 3x3 neighborhood. The spurious ringing pixels are suppressed by applying two thresholds to the *LV*, a low threshold (*Thr_v_low*) and a high threshold (*Thr_v_high*). Since unimpaired pixels exhibit no or very small intensity variance in their neighborhood, a pixel with its *LV* value below or equal to *Thr_v_low* is considered as an unimpaired pixel. In the same way, a pixel with its *LV* value above or equal to *Thr_v_high* is considered as a "noisy pixel". This can be formulated as:

$$VC_n(i,j) = \begin{cases} 1 & Thr\_v\_low < LV(i,j) < Thr\_v\_high \\ 0 & otherwise \end{cases} \tag{7}$$

where

$$LV(i,j) = \frac{1}{9}\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}[I(k,l) - \frac{1}{9}\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}I(k,l)]^2, \qquad i,j \in RR_n \tag{8}$$

$$Thr\_v\_high = \alpha \cdot MAX\,[LV(i,j)], \qquad i,j \in LS_n \tag{9}$$

where VC_n(i, j) indicates the visibility of a ringing pixel at the n[th] ringing region (i.e. RR_n) with its associated line segment (i.e. LS_n) , and LV(i, j) indicates the local variance computed over a 3x3 template, centered at a pixel intensity I(i, j). The value of *Thr_v_low* is chosen to be zero, and the value of *Thr_v_high* is chosen to scale with the strength of corresponding edge (see [23]). Thus, the ringing region RR_n is removed if

$$\frac{SUM\,(VC_n)}{SIZE\,(RR_n)} < R \tag{10}$$

where $SUM(VC_n)$ indicates the number of visible ringing pixels, $SIZE(RR_n)$ indicates the size of the given RR, and R indicates the pre-defined ratio of visible ringing pixels over the detected ringing region.

3.3 The Psychovisual Experiment

To validate our algorithm for ringing region detection, a psychovisual experiment, in which participants were requested to indicate regions of visible ringing in compressed natural images, was carried out. [1] The results were transformed into a subjective ringing region (SRR) map, indicating where in an image on average people see ringing.

3.3.1 Subjective Experiment Procedure

A set of eight source images, reflecting adequate diversity in image content, were taken from the Kodak Lossless True Color Image Suite [30]. Figure 11 shows these source images. They were high resolution and high quality color images of size 768x512 (width × height) pixels. These images were JPEG compressed using MATLAB's *imwrite* function at two different compression levels (i.e. Q=25 and 50). This yielded a test database of sixteen stimuli. These stimuli were displayed on a 17-inch LCD monitor with a screen resolution of 1024x768 pixels. The experiment was conducted in a standard office environment [31] and the viewing distance was approximately 40cm.

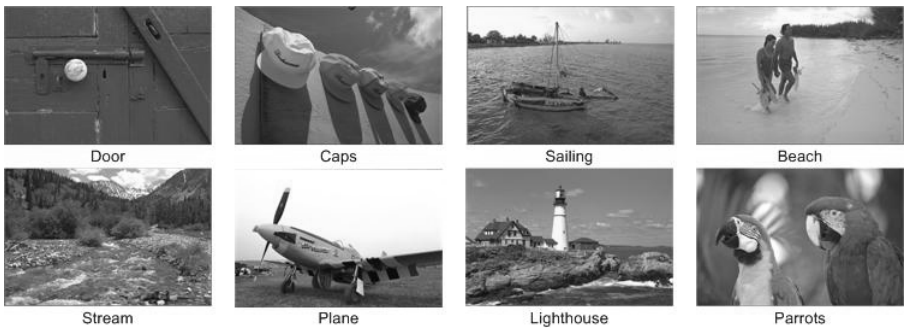


Fig. 11. Source images.

Twelve students of the Delft University of Technology, being eight males and four females, were recruited for the experiment. Before they started the actual assessment, each of them was shown three examples of synthetic ringing, synthetic blocking and synthetic blur artifacts, followed by three real-life images in which ringing, blocking and blur were the most annoying artifacts, respectively. When the participant reported to be able to distinguish ringing from other types of compression artifacts, a set of images with the same level of ringing as used in the rest of the experiment was presented. The participant was requested to mark any region in the image where he/ she perceived ringing, independent of its annoyance. The images used during this training were different from those used in the actual experiment. After training, all 16 stimuli were shown in a random order to each subject in a separate session.

[1] The data collected from this experiment are available to the image quality assessment community on the web-site http://mmi.tudelft.nl/~ingrid/ringing.html

3.3.2 Subjective Data Processing

The recorded edges per image and subject were transformed into a binary image, in which a white pixel indicated perceived ringing and a black pixel referred to absence of visible ringing. This resulted in an individual ringing region (IRR) map per stimulus and subject. These IRR were then averaged over all subjects to a mean ringing region (MRR) map. From the MRR map, the subjective ringing region (SRR) map was derived by simply applying a threshold (i.e. *Thr_srr*) of 0.5, keeping only those edges near which ringing was perceived by half of the subjects. This threshold was introduced to avoid that subjective outliers would strongly affect the performance comparison between various algorithms. Its actual value is further discussed in Section V.

3.4 Performance Evaluation

Our proposed ringing region detection method is validated with respect to the results of the psychovisual experiment, and its performance is compared to existing alternatives in literature. For this performance comparison, we implemented three ringing region detection algorithms recently proposed: (1) region clustering based ringing artifact measure (referred to as RCRM) [23], (2) morphological filtering based ringing artifact measure (referred to as MFRM) [22], and (3) no-reference ringing artifact measure (referred to as NRRM) [21]. In literature, all three methods are proved to be promising in terms of ringing region detection.

3.4.1 Evaluation Criteria

To evaluate the performance of various ringing region detection algorithms we compared the CRR map as calculated for each of the ringing region detection algorithms to the SRR map derived from the psychovisual experiment. These two binary images (i.e. the CRR and SRR map) were compared visually and via a quantitative correlation.

For the visual assessment we produced a comparison map (M_C), which is an RGB color image generated by

$$M_C = \begin{cases} M_C(:,\ :,\ 1) = M_{CRR}\ \&\ [xor(M_{CRR}\ ,\ M_{SRR})] \\ M_C(:,\ :,\ 2) = M_{CRR}\ \&\ M_{SRR} \\ M_C(:,\ :,\ 3) = M_{SRR}\ \&\ [xor(M_{CRR}\ ,\ M_{SRR})] \end{cases} \tag{11}$$

The G (green) channel is assigned to the logical operator AND of the two binary maps, and so, represents the correlated ringing regions. The R (red) and B (blue) channels are assigned to edges occurring only in the CRR map and the SRR map, respectively, and so, represent the uncorrelated ringing regions between both maps. Black regions represent the absence of visible ringing on both maps.

The objective comparison of the CRR map to the SRR map is quantitatively measured by two correlation coefficients, namely $\rho\ 1$ and $\rho\ 2$, defined as follows:

$$\rho_1 = \frac{\sum [M_{CRR} \ \& \ M_{SRR}]}{\sum M_{SRR}} \tag{12}$$

$$\rho_2 = \frac{\sum \{M_{CRR} \ \& \ [xor(M_{CRR} \ , M_{SRR})]\}}{\sum [\sim M_{SRR}]} \tag{13}$$

The numerator of ρ_1 indicates the total number of correlated pixels between the CRR map and SRR map, while the denominator indicates the size of the ringing regions in the SRR map. Thus, ρ_1 quantifies to what extent the subjective ringing regions are detected by the computational models. However, this coefficient by itself is obviously not enough to reflect the detection accuracy of a computational model. A model might be capable of capturing all subjective ringing regions, just by capturing all edges, also those that do not contain visible ringing. These falsely detected ringing regions consequently degrade particularly the efficiency of a subsequent ringing annoyance measurement. The degree of false detections is quantified by ρ_2. Its numerator indicates the size of regions falsely detected by the computational models, and its denominator indicates the size of regions in the SRR map not detected by the human subjects. Evidently, a higher value of ρ_1 combined with a lower value of ρ_2 implies a good detection model.

3.4.2 Model Calibration

Our proposed ringing region detection algorithm uses a number of parameters that need to be tuned to optimal, but at the same time robust performance over different image content. For this tuning, we used five new images (not part of the psychovisual experiment). These images were also JPEG compressed with the MATLAB's *imwrite* function at Q=25 and 50. A few experts in the area of compression artifacts (mainly the authors) indicated the regions in the image with visible ringing. The resulting data were used for optimizing the performance of our ringing region detection algorithm. Robustness over content was evaluated by applying these optimized parameters to the new image content of the psychovisual experiment.

Parameters for the Edge Extraction

This set of parameters includes the standard deviations (i.e. σ_d and σ_r) for the bilateral filter to control the extent of the smoothing effect, and the hysteresis thresholding (i.e. *Thr_high* and *Thr_low*) of the Canny edge detector to trace strong edges while preventing breaking of continuous edges. For the bilateral filter the selection of σ_d and σ_r has been intensively discussed for natural images in [26], and they were set accordingly to σ_d=3 and σ_r=100 in our experiment (see [32] and [32]). For the edge detector Canny sets the *Thr_high* such that a certain percentage (i.e. *p*) of the total amount of pixels is cumulated in the magnitude histogram of the gradient image, and the *Thr_low* as a fixed fraction (i.e. 0.4) of the *Thr_high* [24]. In our implementation, we used a relatively low value of *Thr_high* (i.e. *p*=85%) in order to prevent loosing relevant edges. This may result in irrelevant LSs in the PEM, but

these LSs are later discarded by applying the HVS model. In other words, the choice for the thresholds of the Canny edge detector affect the efficiency of the model rather than its accuracy. Finally, the threshold for the noise removal in the PEM formation was set to 20 pixels. Again, this parameter affects the efficiency rather than the accuracy of the model.

Parameters for Region Definition

This set of parameters determines the width of the EdReg, DeReg, and FeXReg regions. The EdReg representing edge blur is chosen to be equal to the one-pixel thick LS. In case this value is too small, blur pixels can easily be detected as spurious pixels in a ringing region (as described in Section II.C). The width of the DeReg is set as a single-sided support dimension of four pixels, which approximates the maximal extent of ringing that spreads out to a region surrounding an edge in JPEG compression [4]. The actual width of the DeReg may vary depending on the underlying properties of the coding technique, but can be adjusted according to [34]. The width of the FeXReg is empirically selected to be the same as for the DeReg. We experienced that the FeXReg may cross an object boundary or reach another edge, which consequently results in spurious pixels in a detected ringing region. The suppression of these pixels has been discussed in Section II.C.

Parameters for the HVS

This set of parameters includes two essential thresholds, i.e. *Thr_txt* for texture masking and *Thr_lum* for luminance masking. The performance of our algorithm is fairly insensitive to variations of these thresholds within the range of [0.6, 0.95] and [0, 0.8] for *Thr_txt* and *Thr_lum*, respectively. Varying these thresholds within their respective range results in a variation of ρ_1 and ρ_2 over [85%, 95%] and [1%, 3%], respectively. For the final performance evaluation of our model, we set *Thr_txt*=0.9 and *Thr_lum*=0.75.

Parameters for Spurious Ringing Pixel Detection

This set of parameters contains three threshold values (i.e. *Thr_v_low*, *Thr_v_high* (determined by α as shown in equation (9)) and *R*) to further eliminate undesired regions in the CRR map. It should be admitted that this processing step is a fine-tuned optimization to largely remove e.g. the "unimpaired regions" in the CRR map of an uncompressed (or high bit-rate compressed) image, thus making the subsequent calculation of ringing annoyance [16], [17] more efficient. The parameters are determined as *Thr_v_low*=0, α =0.5 and *R*=0.3. *Thr_v_low* and α are set according to experiments and observations reported in [23], while *R* is empirically chosen. *R* is mainly used to speed up the algorithm rather than to improve its accuracy. The inclusion of the detection of spurious ringing pixels hardly affects the overall performance of our model: including or omitting the detection of spurious ringing pixels corresponds to a deviation in ρ_1 and ρ_2 over a range of [-0.5%, +0.5%]. It should, however, be noted that the concept of removing spurious ringing pixels is mainly important for the ringing annoyance estimation,

and hence, these parameters might need to be calibrated again for the subjective data of ringing annoyance [16], [17].

Selected Parameters for Methods from Literature

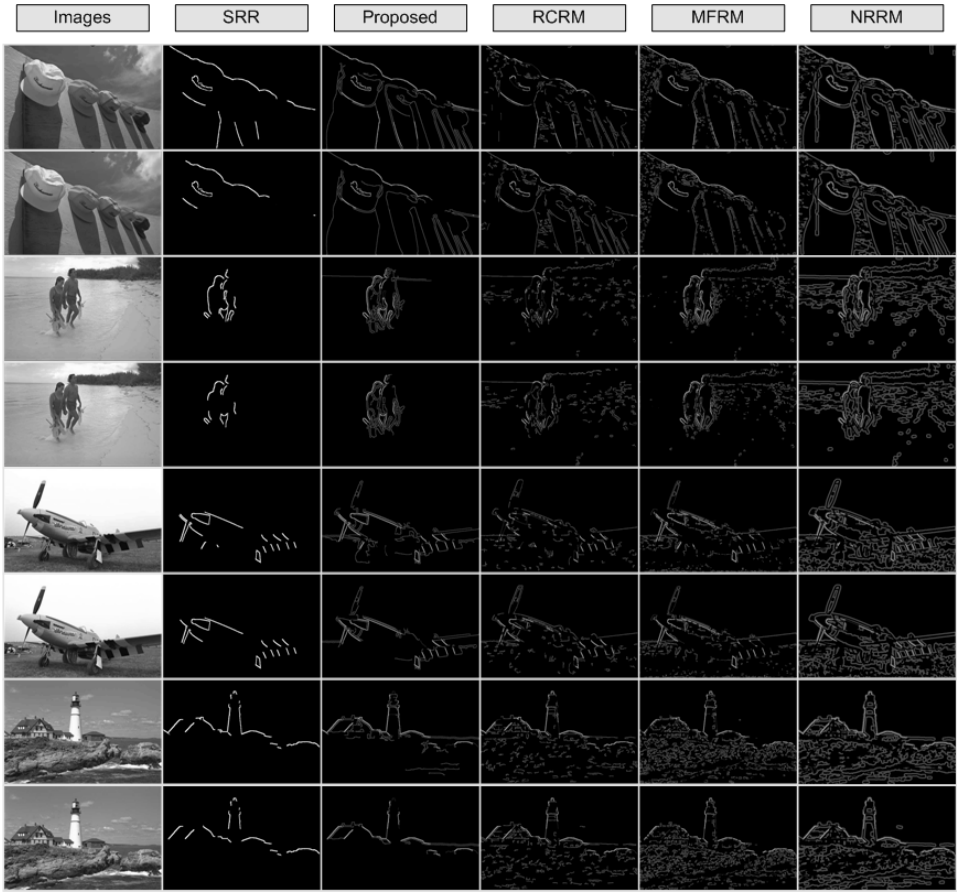
As mentioned above, we will compare the performance of our algorithm to three alternatives published in literature. These methods were implemented following the description in the original publications. However, some important parameters were adjusted to ensure a fair comparison. The parameters to control the thresholding of the edge detector were tuned for each algorithm to yield the highest performance possible for the five test images used during calibration. The parameter for determining the extent of ringing artifacts was equal for all algorithms (i.e. a single-sided ringing region support dimension of 4 pixels).

3.4.3 Evaluation of Overall Model Performance

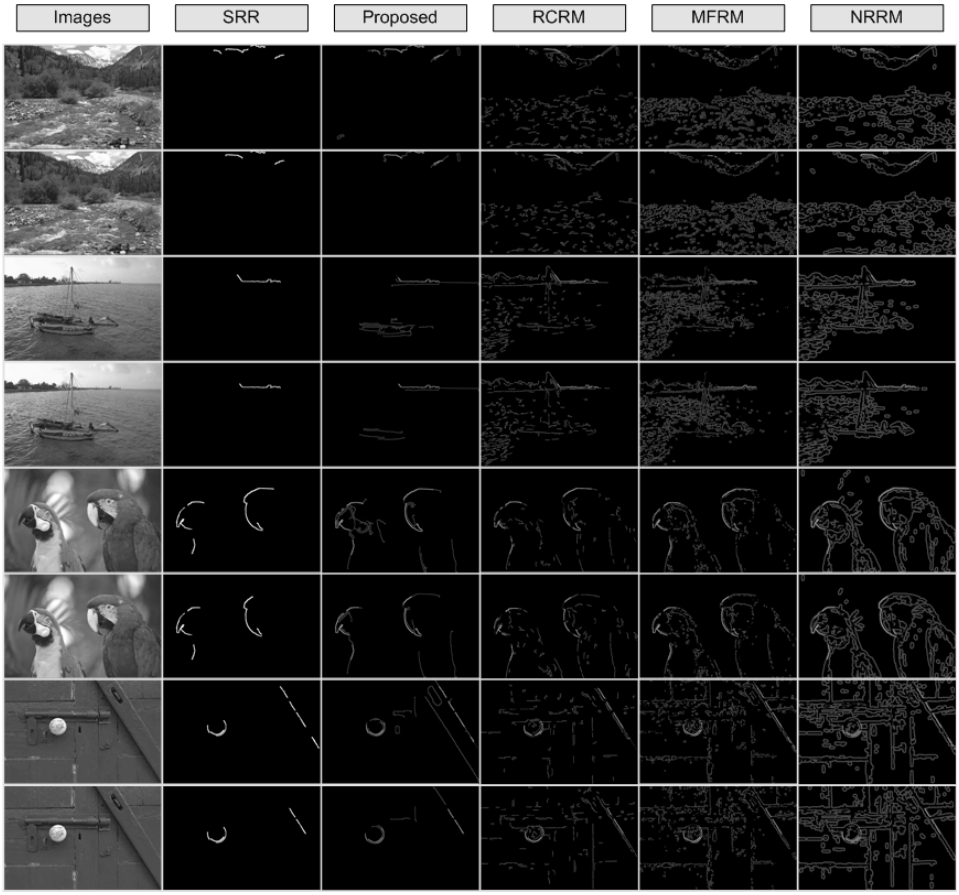
The comparison maps for the visual assessment between the SRR map and the (optimized) CRR maps of the various algorithms are given in Figure 12. The first column shows the test images, the second column presents the SRR maps, and the remaining four columns give the comparison maps of our proposed algorithm, RCRM, MFRM, and NRRM, respectively. In general, most of the ringing regions that were perceived in the psychovisual experiment were also detected by each of the four algorithms. However, our proposed method detects the perceived ringing regions while introducing far less noise (i.e. regions that are not observed subjectively) compared to the other three methods. The correlation coefficients ρ_1 and ρ_2 between the SRR and each of the CRR maps is given in Figure 13. These data are summarized into an overall performance, shown in Table I. In terms of detecting perceived ringing regions (i.e. ρ_1), our proposed method outperforms the other three methods by 15% on average. Also in terms of avoiding false detection (i.e. ρ_2) our method is twice as good as the next best one, namely the RCRM. The latter algorithm, however, is lowest in performance based on ρ_1.

Model	Proposed	RCRM	MFRM	NRRM
$\overline{\rho_1}$	**92%**	**72%**	**79%**	**76%**
$\sigma(\rho_1)$	0.04	0.17	0.20	0.12
$\overline{\rho_2}$	**2.2%**	**4.9%**	**9.6%**	**18%**
$\sigma(\rho_2)$	0.02	0.02	0.04	0.02

Table I. Performance comparison of the four ringing region detection methods (*Thr_srr*=1/ 2 for the SRR maps): mean and standard deviation of the correlation coefficients ρ_1 and ρ_2.

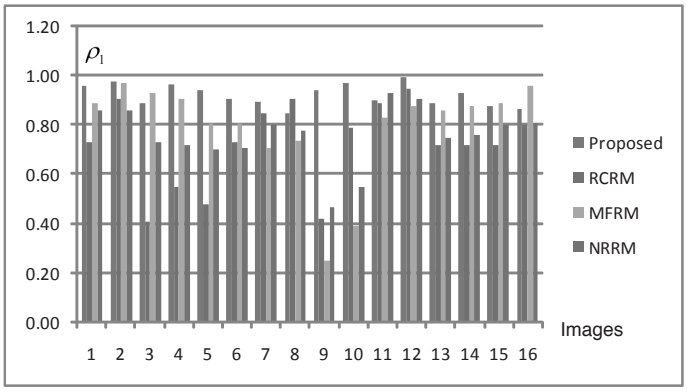


(a)

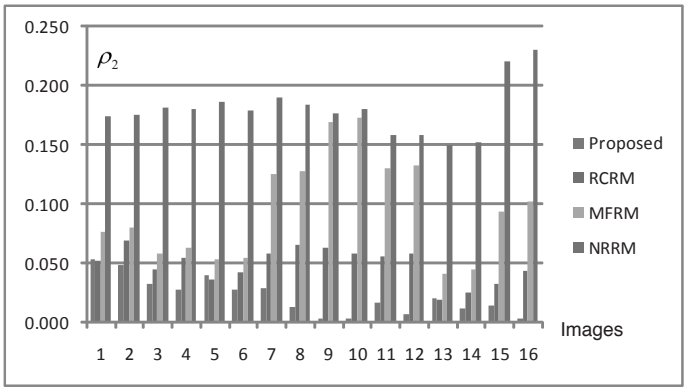


(b)

Fig. 12. Experimental results of visual assessment: (a) Images 1-8: Caps (Q25), Caps (Q50), Beach (Q25), Beach (Q50), Plane (Q25), Plane (Q50), Lighthouse (Q25), Lighthouse (Q50), and (b) Images 9-16: Stream (Q25), Stream (Q50), Sailing (Q25), Sailing (Q50), Parrots (Q25), Parrots (Q50), Door (Q25), Door (Q50). The second column gives the subjective ringing region (SRR) map, and columns 3-6 give the computational ringing region (CRR) map calculated for our proposed approach, the RCRM [23], the MFRM [22], and the NRRM [21], respectively.



(a)



(b)

Fig. 13. Quantitative comparison results: (a) correlation coefficient ρ_1 and (b) correlation coefficient ρ_2.

3.5 Discussion

In this paper we present a novel approach to the detection of regions in an image impaired with visible ringing artifacts. The output of the proposed algorithm serves as input for the second step in the objective ringing metric, existing of the quantification of the actual ringing annoyance in each of the detected regions (as published in [16] and [17]). In this respect it is relevant to realize that a good performance of the ringing region detection algorithm mainly contributes to the efficiency of the second step in the objective metric, rather than to the final accuracy of the prediction in ringing annoyance.

So far, our algorithm is only tested for JPEG compressed image material. More research is needed to also evaluate its performance for different compression techniques. The algorithm is evaluated for two compression levels, and the corresponding CRR maps are highly comparable. Since in this paper we only measure ringing regions, and not ringing annoyance, this is not surprising. Even for

uncompressed images the CRR map will be comparable at first instance, i.e. before removal of spurious ringing regions as discussed in section II.C.

Our proposed ringing region detection algorithm exists of two essential contributions: an edge detector that only preserves perceptually relevant edges and a simple, yet efficient HVS. The use of an ordinary edge detector (as in RCRM, MFRM and NRRM) makes ringing region detection very sensitive to the threshold used; for a high threshold some visually salient edges may not be detected, such that the obvious ringing regions are consequently missed, while for a low threshold many irrelevant edges may be retained, which results in a lot of false ringing regions. Especially for content that is rather insensitive to masking by the HVS (the image "Door" (see Figure 12) is such an example), the number of detected ringing regions strongly depends on the threshold used for the edge detection. The value of our approach is mainly generated by the bilateral filtering (preserving the perceptually relevant edges) rather than by the edge detection itself. The Canny edge detector could have been replaced by a different edge detector, without expected change in performance.

Table I illustrates the advantage of using texture and luminance masking in ringing region detection (as in our proposed method, and in RCRM and MFRM). It obviously reduces the number of detected false ringing regions (lower ρ_2 value). The NRRM, not including HVS properties, clearly has the highest ρ_2 value. From a practical point of view, this may significantly degrade the efficiency, and to some extent the accuracy of predicting ringing annoyance. Including HVS modeling is especially crucial for highly textured images, such as the image "Stream" (see Figure 12). This type of content usually masks ringing to a considerable extent, which should be addressed by a robust HVS model. That our HVS model is sufficiently robust against this demanding content is shown by its highest ρ_1 value and its lowest ρ_2 value compared to the other two algorithms including HVS properties (i.e. RCRM and MFRM, see Figure 13). Additionally, It should be noted that the number of required computations for modeling the HVS is significantly lower for our model than for to the methods RCRM and MFRM. The reduction in complexity is achieved by calculating the HVS only near the perceptually relevant edges and also by simplifying the model of visual masking itself.

The third contribution to our ringing region detection algorithm is a rather ad hoc one: the removal of spurious ringing regions. Due to this spurious ringing region removal, our proposed method captures slightly more visible ringing regions for compression level Q=25 than for compression level Q=50, which is in agreement with the corresponding SRR maps. The impact of compression ratio is less obvious for the other alternative methods. However, this difference in performance is not of major concern, since it can be corrected for in the quantification of actual ringing annoyance, as long as all relevant edges are captured in the ringing region detection. The performance of the ringing region detection algorithms is evaluated against the results of a psychovisual experiment, represented by SRR maps. From the visual assessment in Figure 12, it is clear that all algorithms detect ringing regions that do not occur in the SRR maps. This is not surprising, since the SRR maps are derived such that they only maintain ringing regions detected by most of the participants. Hence, it is possible that some perceptible, but not annoying ringing regions are omitted by applying a threshold to the MRR maps (see Section III.B). To evaluate how the selection of this threshold affects the performance of all algorithms, the

correlation coefficients ρ_1 and ρ_2 are recalculated for a lower threshold of the SRR map (i.e. *Thr_srr*=1/ 3). The results, summarized in Table II, indicate that the actual values of ρ_1 and ρ_2 change for all algorithms, but that the general tendencies are maintained.

Model	Proposed	RCRM	MFRM	NRRM
$\overline{\rho_1}$	**86%**	**74%**	**65%**	**83%**
$\sigma(\rho_1)$	0.09	0.16	0.12	0.12
$\overline{\rho_2}$	**4.5%**	**20%**	**11%**	**31%**
$\sigma(\rho_2)$	0.03	0.08	0.03	0.04

Table II. Performance comparison of the four ringing region detection methods for *Thr_srr*=1/ 3 of the SRR maps: mean and standard deviation of the correlation coefficients ρ_1 and ρ_2.

3.6 Conclusions

In this paper, a novel approach towards the detection of perceived ringing regions in compressed images is presented. The algorithm relies on the compressed image only, which is promising for its applicability in a real-time video chain, e.g. to enhance the quality of artifact impaired video. It adopts a perceptually more meaningful edge detection method for the purpose of ringing region location. This intrinsically avoids the drawback of applying an ordinary edge detector, which has the risk of omitting obvious ringing artifacts near non-detected edges or of increasing the computational cost by measuring ringing visibility near irrelevant edges. The objective detection in agreement with human visual perception of ringing artifacts is ensured by taking into account typical properties of the human visual system, such as texture masking and luminance masking. The human vision model is implemented, based on the local image characteristics around detected edges, to expose only the perceptually prominent ringing regions in an image. The proposed detection method is validated with respect to ringing regions resulting from a psychovisual experiment, and shows to be highly consistent with subjective data. The performance of our approach is compared to existing alternatives in literature, and has been proved to be promising in terms of both reliability and computational efficiency. The proposed ringing region detection method is meanwhile extended with a ringing annoyance metric that can quantify perceived ringing annoyance of compressed images [16], [17].

3.7 References

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

[2] M. Ghanbari, Standard Codecs: Image Compression to Advanced Video Coding. IEE Press, 2003.

[3] I. Richardson, H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia. John Wiley & Sons, 2003.
[4] M. Yuen and H. R. Wu, “A survey of hybrid MC/ DPCM/ DCT video coding distortions,” Signal Processing, vol. 70, no. 3, pp. 247–278, 1998.
[5] M. C. Q. Farias, M. S. Moore, J. M. Foley, and S. K. Mitra, “Perceptual contributions of blocking, blurring, and fuzzy impairments to overall annoyance,” in Proc. of SPIE, Human Vision and Electronic Imaging IX, vol. 5292, 2004, pp. 109–120.
[6] C. C. Koh, S. K. Mitra, J. M. Foley, and I. Heynderickx, “Annoyance of individual artifacts in mpeg-2 compressed video and their relation to overall annoyance,” in Proc. of SPIE, Human Vision and Electronic Imaging X, vol. 5666, 2005, pp. 595–606.
[7] J. Xia, Y. Shi, K. Teunissen, and I. Heynderickx, “Perceivable artifacts in compressed video and their relation to video quality,” Signal Processing: Image Communication, vol. 24, no. 7, pp. 548–556, 2009.
[8] M. Shen and C. J. Kuo, “Review of postprocessing techniques for compression artifact removal,” Journal of Visual Communication and Image Representation, vol. 9, no. 1, pp. 2–14, 1998.
[9] J. Luo, C. Chen, K. Parker, and T. S. Huang, “Artifact reduction in low bit rate dct-based image compression,” IEEE Trans. Image Processing, vol. 5, pp. 1363–1368, 1996.
[10] K. Zon and W. Ali, “Automated video chain optimization,” IEEE Transactions on Consumer Electronics, vol. 47, pp. 593–603, 2001.
[11] Z. Wang, A. C. Bovik, and B. L. Evans, “Blind measurement of blocking artifacts in images,” in Proc. IEEE International Conference on Image Processing, vol. 3, 2000, pp. 981–984.
[12] H. Liu and I. Heynderickx, “A perceptually relevant no-reference blockiness metric based on local image characteristics,” EURASIP Journal on Advances in Signal Processing, vol. 2009, 2009.
[13] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment. available: http:/ / www.vqeg.org,” 2003.
[14] R. Muijs and J. Tegenbosch, “Quality-adaptive sharpness enhancement based on a no-reference blockiness metric,” in Proc. Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2006.
[15] I. O. Kirenko, R. Muijs, and L. Shao, “Coding artifact reduction using non-reference block grid visibility measure,” in Proc. IEEE International Conference on Multimedia and Expo, 2006, pp. 469–472.
[16] H. Liu, N. Klomp, and I. Heynderickx, “A no-reference metric for perceived ringing,” in Proc. Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2009.
[17] H. Liu, N. Klomp, and I. Heynderickx, “A no-reference metric for perceived ringing artifacts in images,” accepted for IEEE Transactions on Circuits and Systems for Video Technology.
[18] A. B. Watson, Digital Image and Human Vision. Cambridge MA: The MIT press, 1993.
[19] B. A. Wandell, Foundations of Vision. Massachusetts: Sinauer Associates, Inc., 1995.

[20] P. Marziliano, F. Dufax, S. Winkler, and T. Ebrahimi, “Perceptual blur and ringing metrics: Application to jpeg2000,” Signal Processing: Image Communication, vol. 19, pp. 163–172, 2004.
[21] R. Barland and A. Saadane, “Reference free quality metric for jpeg-2000 compressed images,” in Proc. International Symposium on Signal Processing and its Applications, vol. 1, 2005, pp. 351–354.
[22] S. H. Oguz, Y. H. Hu, and T. Nguyen, “Image coding ringing artifact reduction using morphological post-filtering,” in Proc. IEEE Second Workshop on Multimedia Signal Processing, 1998, pp. 628–633.
[23] X. Feng and J. Allebach, “Measurement of ringing artifacts in jpegimages,” in Proc. SPIE, vol. 6076, 2006, pp. 74–83.
[24] J. Canny, “A computational approach to edge detection,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 8, no. 6, pp. 679–698, 1986.
[25] R. J. Sternberg, Cognitive Psychology (Fourth Edition). Thomas Wadsworth, 2006.
[26] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in Proc. IEEE International Conference on Computer Vision, 1998, pp. 836–846.
[27] H. Liu, N. Klomp, and I. Heynderickx, “Perceptually relevant ringing region detection method,” in Proc. of the 16th European Signal Processing Conference, 2008.
[28] S. Winkler, “Vision models and quality metrics for image processing applications,” Ph.D. dissertation, 2002.
[29] T. N. Pappas and R. J. Safranek, Perceptual criteria for image quality evaluation. Handbook of Image and Video Processing. Academic Press, 2000.
[30] R. Franzen, “Kodak lossless true color image suite. available: http:/ / www.r0k.us/ graphics/ kodak/ .”
[31] I.-R. R. BT.500-11, Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, Geneva, Switzerland, 2002.
[32] C. Liu, R. Szeliski, S. B. Kang, C. L. Zitnick, and W. T. Freeman, “Automatic estimation and removal of noise from a single image,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 2, pp. 299–314, 2008.
[33] S. Paris and F. Durand, “A fast approximation of the bilateral filter using a signal processing approach,” International Journal of Computer Vision, vol. 81, pp. 24–52, 2009.
[34] S. H. Oguz, “Morphological post-filtering of ringing and lost data concealment in generalized lapped orthogonal transform based image and video coding,” Ph.D. dissertation, 1999.

Chapter 4

A No-Reference Metric for Perceived Ringing Artifacts in Images

***Abstract*: A novel no-reference metric that can automatically quantify ringing annoyance in compressed images is presented. In the first step a recently proposed ringing region detection method extracts the regions which are likely to be impaired by ringing artifacts. To quantify ringing annoyance in these detected regions, the visibility of ringing artifacts is estimated, and is compared to the activity of the corresponding local background. The local annoyance score calculated for each individual ringing region is averaged over all ringing regions to yield a ringing annoyance score for the whole image. A psychovisual experiment is carried out to measure ringing annoyance subjectively and to validate the proposed metric. The performance of our metric is compared to existing alternatives in literature and shows to be highly consistent with subjective data.**

This chapter is based on the research article published as "A No-Reference Metric for Perceived Ringing Artifacts in Images" by H. Liu, N. Klomp and I. Heynderickx in IEEE Transactions on Circuits and Systems for Video Technology, vol. 20, pp 529-539, April, 2010.

4.1 Introduction

Objective metrics have the aim to automatically provide a quantitative measure for image quality aspects, and to eventually serve as computational alternative for expensive image quality assessments by human observers. They are of fundamental importance to a broad range of applications, such as the optimization of digital imaging systems, benchmarking of image and video coding, and quality monitoring and control in displays [1]. They are generally classified into full-reference (FR) metrics and no-reference (NR) metrics, depending on the use of the original image or video. FR metrics are based on measuring the similarity or fidelity between the distorted image and its original version, which is considered as a distortion-free reference. The most widely used FR metrics are mean squared error (MSE) and peak signal-to-noise ratio (PSNR). These metrics, however, have long been criticized for their poor correlation with perceived image quality [1]. A lot of research effort is devoted to the development of FR metrics that can reflect the way human beings perceive image quality [2]. Improved alternatives of FR metrics include e.g. the Structural Similarity (SSIM) Index [3] and the Visual Information Fidelity (VIF) Index [4]. Since FR metrics require the access to the original, which is however not always available in real-world applications, they are usually employed as tools for in-lab testing of image and video processing algorithms. NR metrics instead are more practical because the quality prediction is based on the distorted image only. However, designing NR metrics is still an academic challenge mainly due to the limited understanding of the human visual system (HVS).

In the last decades, considerable progress on the development of NR metrics is made, and some successful methods are reported in the literature [5]-[19]. In [5], natural scene statistics are used to blindly measure the quality of images compressed by JPEG2000. The approach in [5] relies on the assumption that typical natural images exhibit strong statistical regularities, and therefore, reside in a tiny area of the space containing all possible images. Based on this assumption it quantifies image quality by detecting variations in statistical image features in the wavelet domain. In [6] and [7], NR image quality assessment is formulated as a machine learning problem, in which the HVS is treated as a black box whose input-output relationship, such as the one between image characteristics and the quality rating, is to be learned. After appropriate training with subjective data, these models proved to be able to consistently predict the perceived quality of JPEG compressed or otherwise distorted images.

A large number of NR metrics, proposed e.g. in [8]-[19], are based on directly measuring a specific type of artifact created by a specific image distortion process, such as blur caused by acquisition systems, sensor noise, and compression artifacts. In such a scenario, the design of the NR metric can make use of the specific characteristics of the artifact, and therefore, generally obtains a higher reliability with perceived quality degradation [1]. Fortunately, in many practical applications, the distortion processes involved are known, and thus, the design of specific NR metrics turns out to be much more realistic and useful. They can, for example, be combined to predict the *overall* perceived quality. Various examples of this approach are given in literature. A blockiness metric (see e.g. [8]-[11]) can be combined with a flatness metric (see e.g. [12] and [13]) to evaluate the quality of images or video after block-based compression. A ringing metric and a blur metric

are often combined to assess the image quality of wavelet-based compression (see e.g. [14]-[16]). In [17] and [18], multiple artifact metrics are adopted to predict the overall quality of still images or video. In addition to assessing the overall image quality, these specific artifact metrics *individually* are beneficial for optimizing real-time digital imaging systems ([20]-[22]). In the video chain of current TV-sets, various NR metrics, which quantify the quality of the incoming video based on the occurrence of individual artifacts, are used to adapt the parameter settings of the video enhancement algorithms accordingly (see e.g. [23] and [24]). To optimize the performance of both applications mentioned above, reliably modeling specific types of artifacts has clear added value.

Since the widespread use of compression, research on NR metrics is mainly dedicated to compression artifacts and transmission errors [25]. Especially, the blocking artifact, which is one of the most annoying artifacts introduced by block-based compression algorithms [26], such as JPEG or MPEG/ H.263, got a lot of attention. Another compression artifact, especially visible at relatively high bit rates of block-based compression ([21], [26]), but also in wavelet compression [27], is ringing. Unlike the blocking artifact, whose spatial location is very regular and thus easily predictable, the location of ringing is edge dependent, and as such also image content dependent. This makes the task of quantifying ringing annoyance much more difficult. In this paper, we present our recent efforts to develop a NR ringing metric, validate its performance using a subjective study of ringing annoyance in JPEG compressed images, and compare its performance against existing ringing metrics. Before discussing our approach (chapter III) and its performance (chapter IV and V), a more extended explanation of the occurrence and visibility of ringing, and an overview of existing ringing metrics are given in chapter II.

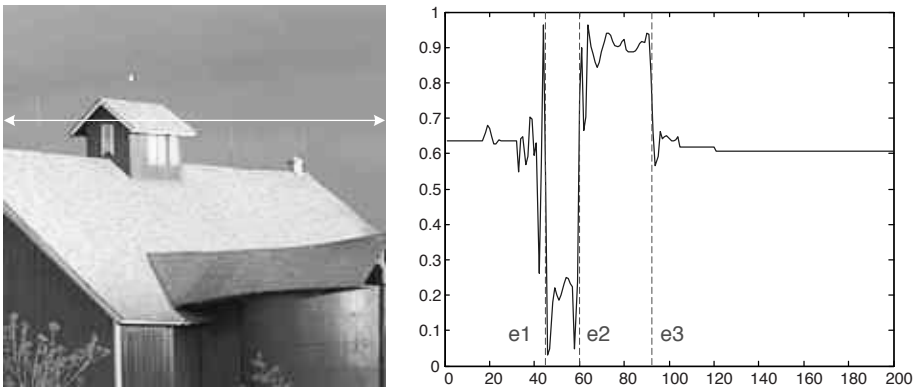
4.2 Background

4.2.1 Perceived Ringing Artifacts

Physical Structure

Current image and video coding techniques are based on lossy data compression, which contains an inherent irreversible information loss. This loss is due to coarse quantization of the image's representation in the frequency domain. The loss within a certain spectral band of the signal in the transform domain reveals itself most prominently at those spatial locations where the contribution from this spectral band to the overall signal power is significant (see [26], [27] and [38]). Since the high frequency components play a significant role in the representation of an edge, coarse quantization in this frequency range (i.e. truncation of the high frequency transform coefficients) consequently results in apparent irregularities around edges in the spatial domain, which are usually referred to as ringing artifacts. More specifically, ringing artifacts manifest themselves in the form of ripples or oscillations around high contrast edges in compressed images. They can range from imperceptible to very annoying, depending on the data source, target bit-rate, or underlying compression scheme [38]. As an example, Figure 1 illustrates ringing artifacts induced by JPEG compression on a natural image.

The occurrence of ringing spreads out to a finite region surrounding the edges, depending on the specific implementation of the coding technique. For example, in DCT coding ringing appears outwards from the edge up to the encompassing block's boundary [26]. An example of how to calculate the extent of the ringing region in a particular codecs is given in [38]. In addition to the edge location dependency, the behavior of ringing also depends on the strength of the edges. It is found in [14], [29] and [38] that, over a wide range of compression ratios, the variance of the ringing artifacts is proportional to the contrast of the associated edge. These important findings have great potential in the design of a reliable ringing metric, and therefore, are explicitly adopted in our algorithm.



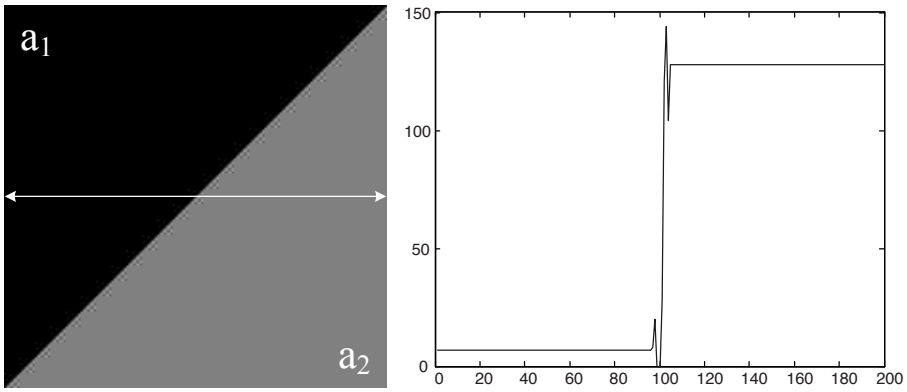
(a) (b)

Fig. 1. Illustration of ringing artifacts: (a) a natural image compressed with JPEG (MATLAB's *imwrite* function with "quality" of 30); and (b) the gray-scale intensity profile along one row of the compressed image (indicated by the solid double arrowhead line in (a)). The dashed lines "e1", "e2" and "e3" indicate the position of the sharp intensity transitions (i.e. edges) along that arrow. Ringing can be perceived as fluctuations in the gray-scale values around the edges at "e1", "e2" and "e3", while the image content here should be uniform.

Masking of the HVS

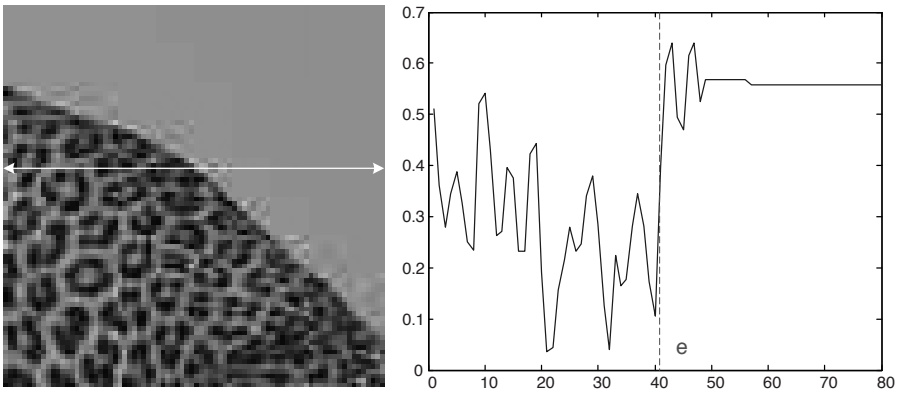
Taking into account the way the HVS perceives artifacts, while removing perceptual redundancies, can be greatly beneficial for matching objective artifact measurement to the human perception of artifacts [39]. Masking designates the reduction in the visibility of one stimulus due to the simultaneous presence of another, and it is strongest when both stimuli have the same or similar frequency, orientation, and location [41]. It is basically due to the limitations in sensitivity of a certain cell or neuron at the retina in relation to the activity of its surrounding cells and neurons. There are two fundamental visual masking effects highly relevant to the perception of ringing artifacts ([28]-[31]). The first one is luminance masking, which refers to the effect that the visibility of a distortion (such as ringing) is maximum for medium background intensity, and it is reduced when the distortion occurs against a very low or very high intensity background [40]. This masking phenomenon happens because of the brightness sensitivity of the HVS, where the

average brightness of the surrounding background alters the visibility threshold of a distortion [42]. The second masking effect is texture masking, which refers to the observation that a distortion (such as ringing) is more visible in homogenous areas than in textured or detailed areas [40]. In textured image regions, small variations in the texture are masked by the macro properties of genuine high frequency details, and therefore, are not perceived by the HVS [38]. The effect of luminance and texture masking on ringing artifacts is illustrated in Figure 2 and Figure 3, respectively.



(a) (b)

Fig. 2. An example of luminance masking on ringing artifacts: (a) an image patch compressed with JPEG (MATLAB's *imwrite* function with "quality" of 30); and (b) the pixel intensity profile along one row of the compressed image patch (indicated by the solid double arrowhead line in (a)). The original image includes two adjacent parts with different gray-scale levels (i.e. 5 for "a1" and 127 for"a2"). Note that although both sides of a step edge exhibit ringing artifacts, the visibility of ringing differs.



(a) (b)

Fig. 3. An example of texture masking on ringing artifacts: (a) an image patch extracted from a JPEG compressed image of bit rate 0.59 bits per pixel (bpp); and (b) the pixel intensity profile along one row of the compressed image patch (indicated by the solid double arrowhead line in (a))., the dashed line "e" indicates the object boundary edge. Note that although both sides of the edge at "e" exhibit ringing artifacts, the visibility of ringing differs.

4.2.2 Existing Ringing Metrics

Until recently, only a limited amount of research effort was devoted to the development of a ringing metric. Some of these metrics are FR, others NR. A FR approach presented in [14] starts from finding important edges in the original image (noise and insignificant edges are removed by applying a threshold to the Sobel gradient image), and then measures ringing around each edge by calculating the difference between the processed image and the reference. Since this metric needs the original image, it has its limitations e.g. for the application in a TV chain. The NR ringing metric, proposed in [17], performs a anisotropic diffusion on the image and measures the noise spectrum filtered out by the anisotropic diffusion process. The basic idea behind this metric is that due to the effectiveness of anisotropic diffusion on deringing, the artifacts would be mostly assimilated into the spectrum of the filtered noise. The NR ringing metric described in [16] indentifies the ringing regions around strong edges in the compressed image, and defines ringing as the ratio of the activity in middle low over middle high frequencies in these ringing regions. An obvious shortcoming of the metrics defined in [14], [16], and [17] is the absence of masking, typically occurring in the HVS, with the consequence that these metrics do not always reflect *perceived* ringing. Typical masking characteristics, such as luminance and texture masking, are explicitly considered in the metrics defined in [28] and [29], in which ringing regions are no longer simply assumed to surround all strong edges in an image, but are determined by a model of the HVS. Including a HVS model in an objective metric might improve its accuracy, but often is computationally intensive for real-time applications. For example, the HVS model used in the metric presented in [28] largely depends on a parameter estimation procedure, which requires a number of calculations to achieve an optimal selection. The model described in [29] is based on a computationally heavy clustering scheme, including both color clustering and texture clustering. From a practical point of view, it is highly desirable to reduce the complexity of the HVS based metric without compromising its overall performance.

The essential idea behind most of the existing metrics mentioned so far (see e.g. in [14], [16] and [28]) is that they consist of a two-step approach. The first step identifies the spatial location, where perceived ringing occurs, and the second step quantifies the visibility or annoyance of ringing in the detected regions. This approach intrinsically avoids the estimation of ringing in irrelevant regions in an image, thus making the quantification of ringing annoyance more reliable, and the calculation more efficient. Additionally, a local determination of the artifact metric provides a spatially varying quality degradation profile within an image, which is useful in e.g. video chain optimization as mentioned in chapter I. Since ringing occurs near sharp edges, where it is not visually masked by local texture or luminance, the detection of ringing regions largely relies on an edge detection method followed by a HVS model. Existing methods (such as e.g. [14], [16], [28] and [29]) usually employ an ordinary edge detector, where a threshold is applied to the gradient image to capture strong edges. Depending on the choice of the threshold, this runs the risk of omitting obvious ringing regions near non-detected edges (e.g. in case of a high threshold) or of increasing the computational cost by modeling the rather complex HVS near irrelevant edges (e.g. in case of a low threshold). This implies that to ensure a reliable detection of perceived ringing while maintaining

low complexity for real-time applications, an efficient approach for both detecting relevant edges and modeling the HVS is needed. Quantification of the annoyance of ringing in the detected areas can be easily achieved by calculating the signal difference between the ringing regions and their corresponding reference, as used in the FR approach described in [14]. However, for a NR ringing metric, the quantification of ringing becomes more challenging mainly due to the lack of a reference. Metrics in literature (such as in [16] and [28]) estimate the visibility of ringing artifacts from the local variance in intensity around each pixel within the detected ringing regions, and average these local variances over all ringing regions to obtain an overall annoyance score. This approach, however, has limited reliability, since it does not include background texture in the ringing regions, which might affect ringing visibility.

To validate the performance of a ringing metric, its predicted quality degradation should be evaluated against subjectively perceived image quality. To prove whether a ringing metric is robust against different compression levels and different image content, the correlation between its objective predictions and subjective ringing ratings must be calculated. Unfortunately, only the performance of the metric reported in [14] is evaluated against subjective data of perceived ringing. For all other metrics (such as the ones in [15], [16], [17] and [28]) nothing can be concluded with respect to their performance in predicting perceived ringing. Since we had no access to the data used in [14] for our metric evaluation, we performed our own subjective experiment.[2]

In this paper, we propose a NR ringing metric based on the same two-step approach mentioned above. For the first step, we rely on our ringing region detection method (see [30] and [31]), the performance of which in terms of extracting regions with perceived ringing has been shown to be promising [31]. Therefore, we consider this part of the metric readily applicable for the second step, in which the ringing annoyance is quantified. To quantify ringing annoyance, we consider each detected ringing region as a perceptual element, in which the local visibility of ringing artifacts is estimated. The contrast in activity between each ringing region and its corresponding background is calculated as the local annoyance score, which is then averaged over all ringing regions to yield an overall ringing annoyance score. It should be noted that the proposed metric is built upon the luminance component of images only in order to reduce the computational load. The performance of the NR metric is evaluated against subjective ringing annoyance in JPEG compression.

4.3 Proposed NR Ringing Metric

4.3.1 Perceived Ringing Region Detection

For the design of our ringing region detection method (see [30] and [31]), we explicitly exploited the specific physical structure of ringing artifacts and some properties of the HVS. The overall proposed algorithm is schematically shown in

[2] The data collected from this experiment are available to the image quality assessment community on the web-site http://mmi.tudelft.nl/~ingrid/ringing.html

Figure 4, which mainly consists of two processing steps: (1) extraction of edges relevant for ringing, which results in a perceptual edge map (PEM), and (2) detection of perceived ringing regions, which yields a computational ringing region (CRR) map. This method is already described in more detail in [30] and [31], and is only briefly repeated here.

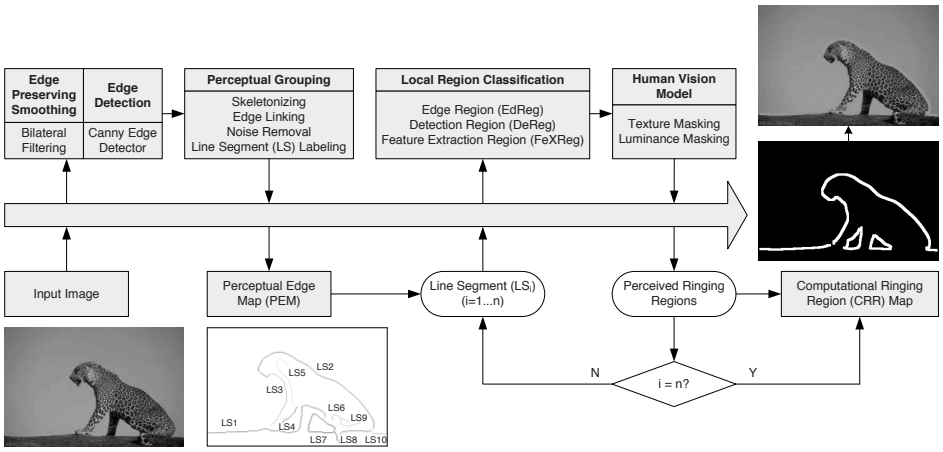


Fig. 4. Schematic overview of the proposed ringing region detection method. In the perceptual edge map (PEM), each perceptually relevant line segment (LS) is labeled in a different color. In the computational ringing region (CRR) map, the white areas indicate the detected perceived ringing regions, and the spatial location of these regions is illustrated in a separate image by green areas.

To extract the most relevant edges for the purpose of ringing detection, an advanced edge detector is used. It adopts a bilateral filter [32] to largely smooth "irrelevant edges" (i.e. in textured areas), while the position of the "relevant edges" (e.g. contours of objects) is retained. Subsequently, a Canny edge detector [33] is applied on the filtered image to obtain the "relevant edges". The detected edges are combined into line segments (hereafter referred to as LS), which are defined as elements of connected edge pixels. These LSs are constructed over the Canny edge map by a simple grouping process, including skeletonizing, edge linking, noise removal and LS labeling. Figure 4 shows the extracted PEM, which is formed by a set of these LSs. It clearly illustrates the selection of the edges more relevant for ringing (i.e. the contours of the leopard) in combination with the avoidance of the irrelevant edges (i.e. the texture in the skin of the leopard).

To select the edges around which ringing is actually perceived each LS of the PEM is examined individually on the occurrence of perceived ringing. To this end, the region around a LS is divided into three zones: the edge region (i.e. EdReg), the detection region (i.e. DeReg) and the feature extraction region (i.e. FeXReg). First, the level of texture or detail is estimated from the FeXReg, and those parts of the DeReg, in which the visibility of ringing is masked by texture, are discarded. Subsequently, the average luminance in each remaining part of the DeReg is calculated and those parts with a value above or below a certain threshold are discarded. In this way, only those regions around each LS, in which ringing is

visible, are extracted, and then accumulated in the CRR map as illustrated in Figure 4.

4.3.2 Ringing Annoyance Estimation

The CRR map indicates the spatial location of perceived ringing, but it does not give any information yet on how annoying the ringing artifacts in the detected region are. To quantify ringing annoyance, we first split up the detected region in the CRR map into so-called ringing objects (ROs). Figure 5 illustrates the definition of a RO. It starts from the LSs of the PEM, shown in Figure 4. Each LS is considered to be split up in a set of connected components (i.e. objects) depending on the local level of texture and averaged luminance in its DeReg (as defined in [30] and [31]). Then, by using the model of the HVS, the visibility of ringing in each object is determined. By removing the objects, in which ringing is invisible due to masking, the remaining objects are defined as ROs. As an example, illustrated in Figure 5(b) the LS1 of the PEM in Figure 4 is split up in two ROs, while the LS2 remains as one RO. Some of the LSs, e.g. LS5, LS6, LS8 and LS9, do not result in a RO, since no visible ringing is detected around this LS based on the HVS. So, each RO intrinsically is a single cluster resulting from the application of the human vision model to the LSs of the PEM. Hence, the definition of a RO fully relies on the local image content, and as such, is independent of scaling or cropping the image. Once the ROs are defined (as illustrated in Figure 5(c)), a ringing annoyance score (RAS) is calculated for each of them, and the overall annoyance score for the image is simply the mean of the RAS over all ROs.

The approach taken to quantify perceived ringing is inspired by the basic idea used in the FR metric [14], and is accomplished by the following two steps: (1) calculating the activity of each RO; and (2) comparing that activity to the activity in the neighboring background to which the RO belongs.

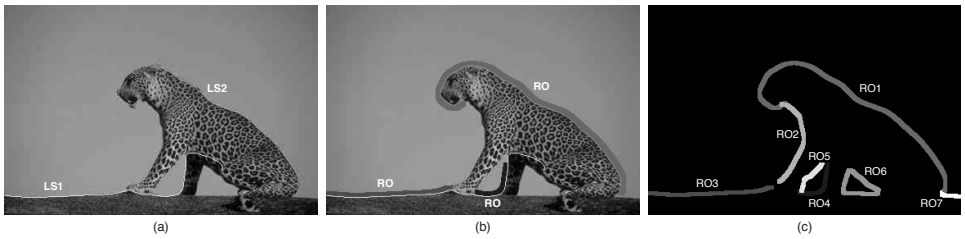


Fig. 5. Illustration of the definition of a ringing object (RO): (a) original JPEG image and two (out of ten) of its detected line segments (LSs) (i.e. LS1 and LS2 of the PEM in Figure 4), (b) implementation of the human vision model to LS1 and LS2, resulting in two separate ROs for LS1 and one RO for LS2, and (c) all detected ROs as a result of applying the human vision model to the whole PEM (i.e. ten LSs); they are indicated with different colors.

Region Assignment

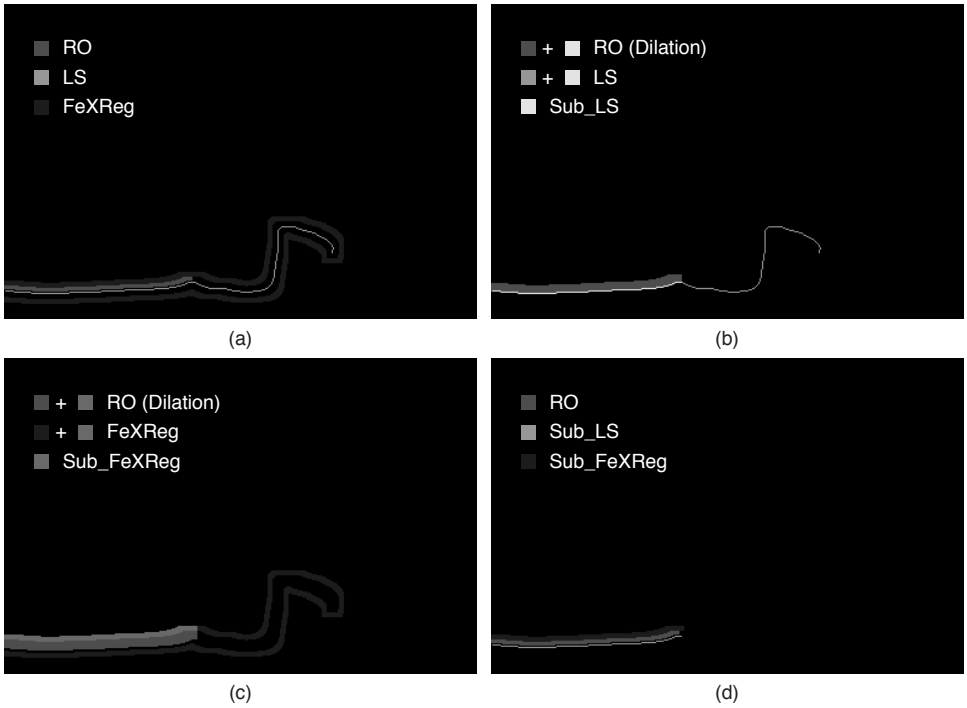


Fig. 6. Illustration of region assignment: (a) a ringing object (RO) (see "RO3" in Figure 5(b)) with its corresponding line segment (LS) and feature extraction region (FeXReg), (b) the corresponding edge of LS covered by the dilated RO is assigned as the Sub-LS, (c) the corresponding region of FeXReg covered by the dilated RO is assigned as the Sub-FeXReg, and (d) the results of region assignment.

To implement the two steps mentioned above, we first assign two relevant components to each RO in the CRR map: (1) the edge corresponding to each LS (i.e. referred to as Sub-LS), which is used to determine whether a pixel in the RO is a visible ringing pixel, and (2) the corresponding FeXReg region (i.e. referred to as Sub-FeXReg), which is employed as the reference for the RO. The FeXReg is located far away from the LS, and thus unlikely to be impaired by ringing artifacts. This region assignment is implemented by thickening a RO with a dilation operation. The corresponding LS and FeXReg which are covered by the RO during the dilation process are referred to as the Sub-LS and Sub-FeXReg, respectively. Figure 6 illustrates this procedure. A specific RO (i.e. "RO3" in the CRR map of Figure 5) with its corresponding LS and FeXReg are shown in Figure 6(a). When dilating the RO with a square structuring element of 5 pixels width (e.g. for an image of 256x384 (height x width) pixels), the region of LS which is covered by the expanded RO is assigned as the Sub-LS (i.e. the yellow region in Figure 6(b)). The Sub-FeXReg (i.e. the purple region in Figure 6(c)) is assigned in the same way by dilating the RO with a square structuring element of 9 pixels width. The resulting Sub-LS and Sub-FeXReg are shown in Figure 6(d). It is noted that the size of the structuring element should be linearly scaled with the image size. The region assignment mentioned above is performed for each RO in the CRR map to eventually obtain a list of coordinates, which indicates the spatial location of each individual RO and its

corresponding Sub-LS and Sub-FeXReg. Figure 7 indicates the format of such a resulting list of coordinates. This way of working intrinsically facilitates the subsequent local analysis and processing of image characteristics.

$$\begin{bmatrix} \text{RO1} \\ \text{RO2} \\ \text{RO3} \\ \cdots \\ \text{ROn} \end{bmatrix} = \begin{bmatrix} \textit{Coord}\{\text{RO1}\} ; \textit{Coord}\{\text{Sub-LS1}\} ; \textit{Coord}\{\text{Sub-FeXReg1}\} \\ \textit{Coord}\{\text{RO2}\} ; \textit{Coord}\{\text{Sub-LS2}\} ; \textit{Coord}\{\text{Sub-FeXReg2}\} \\ \textit{Coord}\{\text{RO3}\} ; \textit{Coord}\{\text{Sub-LS3}\} ; \textit{Coord}\{\text{Sub-FeXReg3}\} \\ \cdots \\ \textit{Coord}\{\text{ROn}\} ; \textit{Coord}\{\text{Sub-LSn}\} ; \textit{Coord}\{\text{Sub-FeXRegn}\} \end{bmatrix}$$

Fig. 7. Illustration of the list of coordinates as the result of region assignment (the total number of ringing objects (RO) in the CRR map is n).

Local Visibility of Ringing Pixels

Since ringing manifests itself in the form of artificial oscillations in the spatial domain, its local behavior can be reasonably described as the intensity variance of pixels in the neighborhood [28], [29]. In this paper, determining whether a pixel in a RO is a visible ringing pixel is based on calculating the local variance (LV) in intensity in its 3x3 neighborhood, which is formulated as

$$LV(i,j) = \frac{1}{9}\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}[I(k,l) - \frac{1}{9}\sum_{k=i-1}^{i+1}\sum_{l=j-1}^{j+1}I(k,l)]^2 , \quad (i,j) \in Coord\{RO_n\} \qquad (1)$$

where LV(i, j) denotes the local variance computed over a 3x3 template, centered at pixel (i, j) having an intensity I(i, j) within the nth ringing object (i.e. RO_n).

The LV only yields an accurate result in case the RO is originally smooth around the edge; indeed, otherwise the LV can be high due to the activity of a textured or edge pixel. One would expect that the issue of considering texture as ringing is efficiently avoided by the application of a texture masking model in the ringing region detection phase (see [30] and [31]). However, we experienced that the dilation operation used in the human vision model may misclassify certain edge or texture components into a RO. In addition, there might be pixels in the RO exhibiting no or a very small intensity variance in their neighborhoods, which means they are not impaired by ringing artifacts (e.g. in higher bit-rate compression). This implies that a RO still possibly contains spurious ringing pixels, which manifest themselves either as "noisy pixels" (i.e. misclassified edge or texture pixels) or as "unimpaired pixels" (i.e. pixels with a very low variance in intensity in the neighborhood). Figure 8 gives an example of the image content underneath a detected RO (i.e. "RO2" as illustrated in Figure 5), where noisy pixels and unimpaired pixels coexist with real ringing pixels. Calculating the LV over these spurious ringing pixels may degrade the accuracy of measuring the actual ringing activity. The effect of the spurious ringing pixels on the RAS is avoided by applying two thresholds, a high threshold (*Thr_vc_high*) and a low threshold (*Thr_vc_low*). A pixel with its LV value above or equal to *Thr_vc_high* is considered as a "noisy pixel", and its visibility is set to "0". A pixel with its LV value below or equal to

Thr_vc_low is considered as an "unimpaired pixel", and its visibility is also set to "0". Hence,

$$VC(i,j) = \begin{cases} LV(i,j) & Thr\_vc\_low < LV(i,j) < Thr\_vc\_high \\ 0 & otherwise \end{cases} \quad (2)$$

where VC(i, j) indicates the visibility coefficient at location (i, j) within the RO_n. After parameter optimization the value of *Thr_vc_low* is chosen to be zero, and the value of *Thr_vc_high* is chosen to scale with the strength of the corresponding edge, since we found that the actual LV range corresponding to a visible ringing pixel depends on the strength of its corresponding Sub-LS. Thus, *Thr_vc_high* is defined as

$$Thr\_vc\_high = \alpha \cdot MAX\ [LV(i,j)], \qquad (i,j) \in Coord\{Sub\_LS_n\} \quad (3)$$

where LV is calculated over the Sub-LS (i.e. Sub_LS_n) assigned to the RO_n, and α (specified in Section V) is used to adjust the value of the high threshold.

All visible ringing pixels are extracted from each individual RO, and their visibility is indicated by a visibility coefficient (VC) according to (2). Figure 9 illustrates the extraction of visible ringing pixels in an image, in which their visibility is indicated by a different color in a color bar.

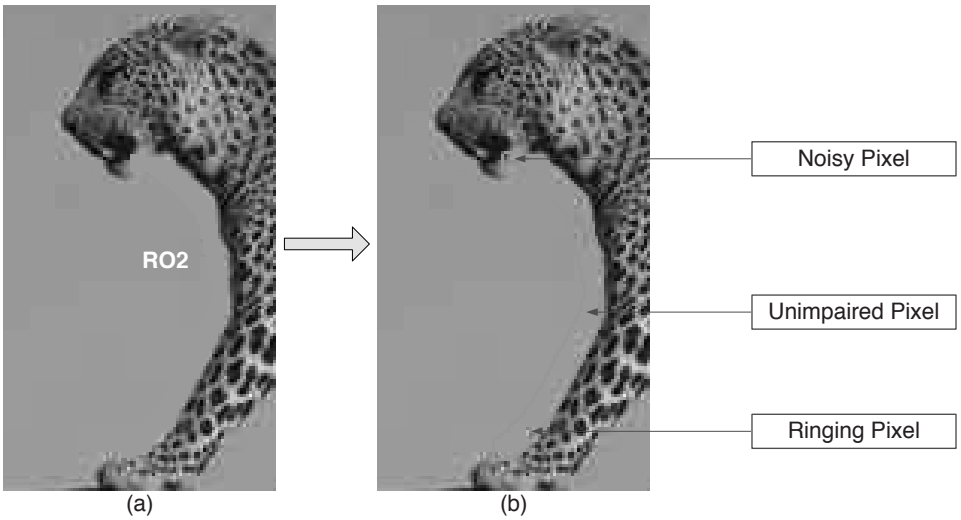


Fig. 8. Illustration of three types of pixels within a ringing object (RO): (a) original JPEG image and a detected RO (see "RO2" in Figure 5(b)), and (b) illustration of the image content underneath the corresponding RO, in which noisy pixels and unimpaired pixels coexist with real ringing pixels.

Fig. 9. Illustration of visible ringing pixel extraction: (a) original JPEG image, and (b) extracted visible ringing pixels with their visibility indicated with a color ranging from black (no visibility) to yellow for the highest visibility.

Ringing Annoyance Estimation

The visibility coefficient for each ringing pixel in itself is yet insufficient to reflect the way human beings perceive ringing. It is the contrast between the visibility of a ringing artifact and its corresponding background that causes the perception of ringing annoyance [29], [38]. More strongly visible ringing pixels against a smoother background are most annoying. Since the Sub-FeXReg is already assigned to each RO to represent its local background, the activity of the Sub-FeXReg is readily calculated as the mean local variance (MLV)

$$MLV(Sub-FeX\,\mathrm{Re}\,g_n)=\frac{1}{Ns}\sum LV(i,j),\quad (i,j)\in Coord\{Sub-FeX\,\mathrm{Re}\,g_n\} \qquad (4)$$

where Ns indicates the total number of pixels within the nth Sub-FeXReg (i.e. $Sub\text{-}FeXReg_n$), and LV(i, j) indicates the local variance calculated at pixel location (i, j) within the $Sub\text{-}FeXReg_n$. For the corresponding RO (i.e. RO_n), its activity is defined as:

$$MLV(RO_n)=\frac{1}{Nr}\sum VC(i,j),\quad (i,j)\in Coord\{RO_n\}\ \&\ VC(i,j)\neq 0 \qquad (5)$$

where Nr indicates the total number of visible ringing pixels within the RO_n, and VC(i, j) indicates the visibility coefficient (see (2)) calculated at pixel location (i, j) within the RO_n.

Once the activity of a RO and of its corresponding Sub-FeXReg is calculated, the difference between them is used to quantify the ringing annoyance for this RO. Hence, the ringing annoyance score (RAS) is defined as:

$$RAS(RO_n)=No\times[MLV(RO_n)-MLV(Sub\_FeX\,\mathrm{Re}\,g_n)] \qquad (6)$$

where No indicates the total number of pixels within the RO_n.

Based on the annoyance score per RO the overall ringing annoyance score for an image is calculated according to the procedure schematically shown in Figure 10. It contains removal of ROs, for which the amount of visible ringing pixels is below a threshold R. In our algorithm, R is set as a pre-defined percentage (specified in Section V) of the total number of pixels in the RO. This is done with the estimation accuracy and speed in mind, since these ROs contain a too small number of visible ringing pixels to contribute to the overall perception of ringing annoyance. Eventually, the proposed ringing metric is defined as the mean of the ringing annoyance scores (MRAS) over all remaining ROs, which is formulated as:

$$MRAS = \frac{1}{T}\sum_{n-1}^{N} RAS(RO_n) \tag{7}$$

where N indicates the total number of ROs, excluding the discarded ones, and T indicates the total number of pixels within these N ROs.

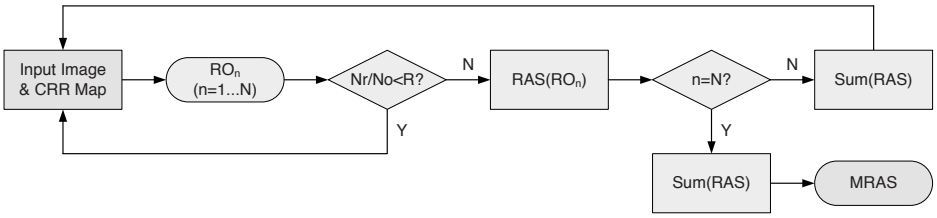


Fig. 10. Schematic overview of the proposed ringing metric (i.e. MRAS).

4.4 Psychovisual Experiment

To validate our proposed ringing metric, a subjective experiment was carried out, in which participants scored the annoyance of ringing artifacts in compressed images.

4.4.1 Experimental Procedure

Image Database and Test Environment

A set of eleven source images, reflecting adequate diversity in image content, was taken from the "Kodak Lossless True Color Image Suite" [35]. Figure 11 shows these source images. They were high resolution and high quality color images of size 768x512 (width x height) pixels. Some images have high activity, while others are mostly smooth. These images were JPEG compressed at four different compression levels (i.e. quality Q=25, 40, 55, 70) using MATLAB's *imwrite* function. This yielded a test database of fifty-five stimuli (including the originals). The compression level was varied over such a range of quality levels that images with a broad range of ringing annoyance, from imperceptible to high levels of impairment, were generated. The stimuli were displayed on a Philips Cineos 37" LCD screen with a native resolution of 1920x1080 pixels and a screen refresh rate of 60 Hz. The experiment was conducted in a standard office environment [34] and the viewing distance was approximately 60cm.

Fig. 11. Source images used in the subjective quality study.

Test Methodology

A single-stimulus (SS) method was used in our experiment, which means that subjects had to score the ringing annoyance for each stimulus in the absence of a reference. The scoring scale ranged from 0 to 100, where "0" means no ringing annoyance and "100" means highest ringing annoyance. The quality scale included additional semantic labels (i.e. "low", "average", and "high" ringing annoyance) at intermediate points for reference as illustrated in Figure 12.

The participants of the study were recruited from the MSc program of the Department of Mediamatics at the Delft University of Technology. The twenty students, being fourteen males and six females, were inexperienced with image quality assessment and coding artifacts. Before the start of the experiment, an instruction about the goal and procedure (e.g. the type of assessment, the scoring scale and the timing) of the experiment was given to each individual subject. A training session was conducted showing three examples of synthetic ringing, synthetic blocking and synthetic blur, followed by three real-life images in which ringing, blocking and blur were the most annoying artifacts, respectively. When the subject reported to understand ringing and to be able to distinguish it from other types of compression artifacts, a set of ten images covering the same range of ringing annoyance as used in the actual study was presented to the subject in order to familiarize him or her with how to use the range of the scoring scale. Then, three stimuli were shown one by one and the participant exercised how to indicate ringing annoyance on the scoring scale. The images used in the training session were different from those used in the actual experiment. After training, the test images were shown in a random order to each subject in a separate session.

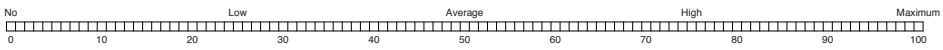


Fig. 12. Quality scale used in the ringing annoyance assessment.

4.4.2 Processing of the Raw Data

Outlier Detection and Subject Rejection

Before the actual data analysis, a simple outlier detection and subject rejection model was implemented on the raw annoyance scores. An individual score for an image was considered to be an outlier if it was outside an interval of two standard deviations around the mean score for that image. All annoyance scores of a subject were rejected if more than five of his/ her scores were outliers. Overall, one subject out of twenty was rejected, and about 3% of the scores were rejected as outliers.

MOS Scores

After outlier removal and subject rejection, the scores of the remaining subjects were calibrated using z-scores [36]:

$$z_{ij} = \frac{r_{ij} - \mu_i}{\sigma_i} \tag{8}$$

where r_{ij} and z_{ij} indicate the raw score and z-score for the *i*-th subject and *j*-th image, respectively. μ_i is the mean of the raw scores over all images scored by subject *i*, and σ_i is the corresponding standard deviation. The z-scores were then averaged across subjects to yield a mean opinion score (MOS) for the *j*-th image

$$MOS_j = \frac{1}{S}\sum_{i=1}^{S} z_{ij} \tag{9}$$

where S is the total number of subjects (after subject rejection).

4.5 Performance Evaluation

Our proposed ringing metric is validated with respect to the data resulting from the psychovisual experiment, and its performance is compared to three alternatives recently published in literature: one FR ringing metric, which is referred to as FRRM [14]; and two NR ringing metrics, which are referred to as NRRM [16] and VRM [28], respectively. In literature, these metrics are all proved to be promising in measuring ringing artifacts in compressed images. It should be noted that we implemented these three metrics ourselves based on the information available in the papers and tuned their parameters to yield the highest performance possible for the set of test images used in our experiments. This is done to ensure a fair comparison between the results from different metrics. The parameters used for our proposed metric are specified as follows: (1) for the ringing region detection: σ_d=3 and σ_r=100 for the bilateral filter, *threshold_high*=0.85 and *threshold_low*=0.4 for the Canny edge detector, *Thr_txt*=0.9 and *Thr_lum*=0.75 for the human vision model, and the *EdReg*, *DeReg* and *FeXReg* are determined with a square structuring element whose width is 3, 9 and 17, respectively (see [30] and [31]); and (2) for the ringing annoyance estimation: *Thr_vc_low*=0, α =0.5 and *R*=0.75. It should be noted that

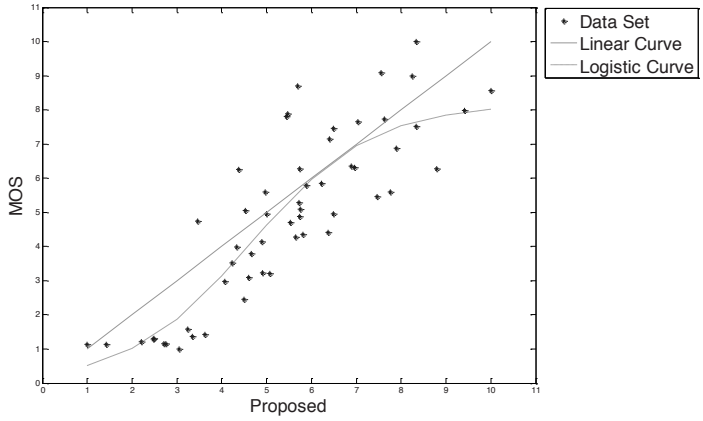
these parameter settings are empirically determined. The first set of parameters for the ringing region detection was defined based on subjective data for ringing region visibility (see [30] and [31]), and is proved in this paper to be robust for a new set of images. The remaining parameters used for the ringing annoyance estimation are determined in pilot experiments on both synthetic patterns and natural images. The performance of the metric is fairly insensitive to variations in the range of [0 0.1] and [0.4, 0.6] for the values of *Thr_vc_low* and α , respectively. The parameter R is mainly used to speed up the algorithm, and thus, hardly affects the prediction accuracy of the metric.

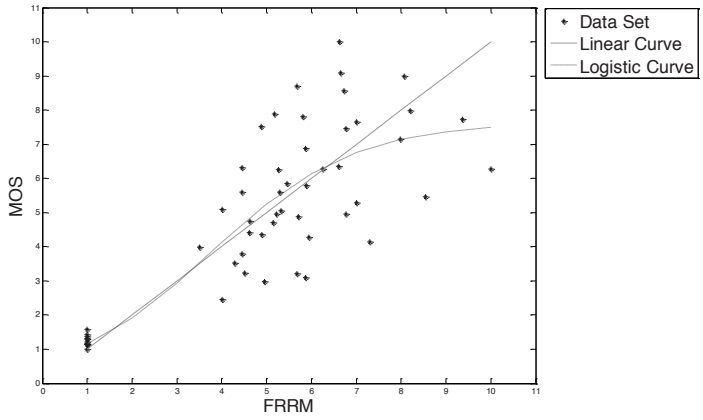
4.5.1 Evaluation Criteria

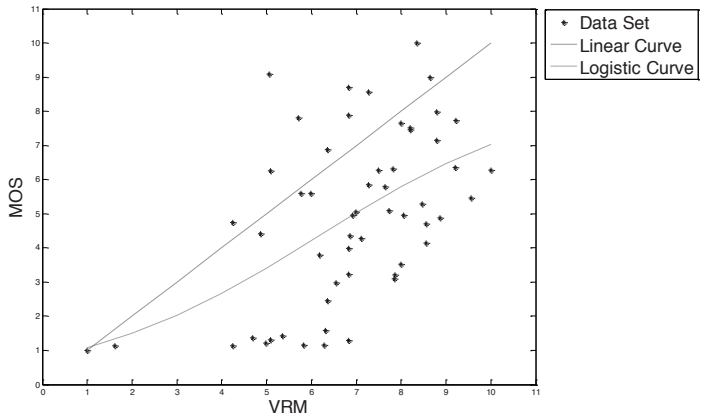
As prescribed by the VQEG [25] the performance of an objective metric can be quantitatively evaluated with respect to its ability to predict subjective quality ratings (the MOS), based on the Pearson linear correlation coefficient to indicate prediction accuracy, the Spearman rank order correlation coefficient to indicate prediction monotonicity, and the outlier ratio to indicate prediction consistency. As suggested in [39], the metric's performance can also be evaluated with non-linear correlations using a non-linear mapping function for the objective predictions before computing the correlation. For example, a logistic function may be applied to the objective metric results to account for a possible saturation effect. A non-linear fitting usually yields higher correlation coefficients in absolute terms, while generally keeping the relative differences between the metrics [39]. On the other hand, without a sophisticated non-linear fitting (often including various parameters) the correlation coefficients cannot mask a bad performance of the metric itself. To better visualize differences in performance we propose to avoid any non-linear fitting and to directly use linear correlation between the metric's predictions and the subjective data. However, to demonstrate the effect of a non-linear mapping, both the linear and non-linear correlations are given in this paper.

4.5.2 Experimental Results

Our proposed ringing metric and the three alternative metrics (i.e. FRRM, NRRM and VRM) are applied to our database of 55 stimuli. Figure 13 shows the scatter plots of the MOS versus our proposed metric, FRRM, NRRM and VRM, respectively. Table I lists the correlation coefficients. To also show the non-linear correlation, a four-parameter logistic function suggested in [25] was used to fit the metric's predictions to the MOS. The resulting curve fits are included in Figure 13, and the correlation coefficients are listed in Table II.


Data Set
Linear Curve
Logistic Curve
MOS
Proposed


Data Set
Linear Curve
Logistic Curve
MOS
FRRM


Data Set
Linear Curve
Logistic Curve
MOS
VRM

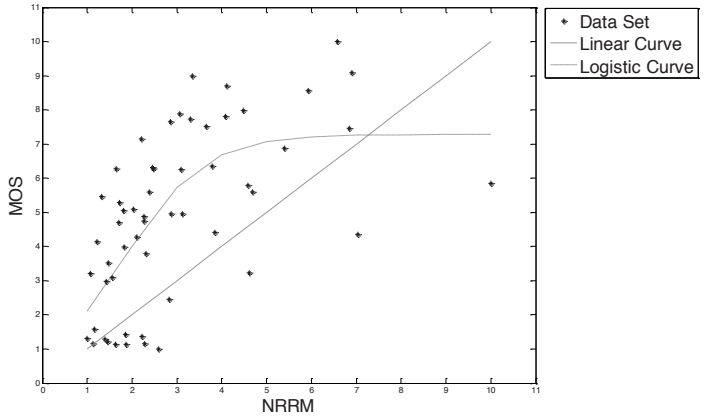


Fig. 13. Scatter plots of MOS vs. the ringing metrics FRRM [14], VRM [28], NRRM [16] and our proposed metric. The full-line curves show the linear fit between the MOS and metric results, while the dashed-line curves show the non-linear logistic fit.

Metric	Pearson Linear Correlation	Spearman Rank Order Correlation	Outlier Ratio
Proposed	**0.851**	**0.850**	**0**
FRRM	**0.793**	**0.744**	**0**
VRM	**0.519**	**0.498**	**0.291**
NRRM	**0.561**	**0.649**	**0.218**

Table. I. Performance comparison of four ringing metrics (our proposed metric, FRRM [14], VRM [28], and NRRM [16]) without non-linear fitting; the threshold to determine the outlier ratio is set to 1.5 standard deviations of the MOS [25].

Metric	Pearson Linear Correlation	Spearman Rank Order Correlation	Outlier Ratio
Proposed	**0.868**	**0.850**	**0**
FRRM	**0.824**	**0.744**	**0.127**
VRM	**0.521**	**0.498**	**0.218**
NRRM	**0.667**	**0.649**	**0.146**

Table. II. Performance comparison of four ringing metrics (our proposed metric, FRRM [14], VRM [28], and NRRM [16]) after a logistic fit of the metrics' predictions to the MOS; the threshold to determine the outlier ratio is set to 1.5 standard deviations of the MOS scores [25].

Figure 13 and Table I demonstrate that our proposed NR ringing metric outperforms the existing metrics in the prediction of ringing annoyance. In comparison to the FR ringing metric FRRM our metric shows a higher correlation to the subjective data, i.e. the gain in the Pearson correlation coefficient is $\Delta P=5\%$, and in the Spearman correlation coefficient is $\Delta S=11\%$. The lower correlation for the FRRM compared to our metric most probably is due to the absence of a HVS model in the FRRM. It simply assumes that ringing occurs unconditionally in regions

surrounding strong edges in an image, neglecting possible luminance and texture masking effects. As a consequence, measuring ringing annoyance in the regions where ringing is invisible to the human eye potentially degrades the prediction performance of this metric. Our metric does contain a model for visual masking, and so, intrinsically avoids the estimation of ringing in irrelevant regions (e.g. texture areas) in an image, thus making the quantification of ringing annoyance more accurate.

Compared to the alternative NR ringing metrics, our metric manifests a much higher prediction performance relative to VRM and NRRM. The measured gain of our metric compared to VRM is $\Delta P=33\%$ and $\Delta S=35\%$, and compared to NRRM is $\Delta P=29\%$ and $\Delta S=20\%$. A possible reason for the lower performance of the NRRM is that it does not take into account spatial masking by the HVS, thus inevitably measuring ringing in some textured regions. Actually the metric may misclassify texture components into ringing artifacts, which may heavily degrade the prediction accuracy of a ringing metric. Comparing the performance of NRRM to that of FRRM (both without a masking model), it is clear that a NR metric is more sensitive to misclassified textured regions than a FR metric. A FR approach can account for the texture by comparing the region to the same unimpaired, but textured region in the reference. As a result, the error of misclassifying texture as ringing is expected (and confirmed) to be smaller.

It should be noted that exactly the same conclusions can be drawn from Table II as discussed above for Table I. This confirms the statement already reported in [39] that non-linear mapping of the metric's predictions to the MOS affects the absolute values, but not the relative differences between metrics.

4.6 Discussion

The experimental results tend to validate our approach in the design a no-reference ringing metric, existing of: (1) a reliable ringing region detection model and (2) a refined ringing annoyance estimation method. The importance of a reliable ringing region detection method can be seen by comparing the metric VRM to the one reported in [37] (which is a previous version of the one reported here, not including yet the comparison of the variance with the background and the detection of spurious ringing pixels). In both metrics, the annoyance score is simply defined as the intensity variance in the detected ringing regions. The only difference between them lies in the HVS model included in the metric of [37] for detecting *perceived* ringing regions. Therefore, the performance gain of the metric of [37] (with a Pearson correlation coefficient of **0.8**) over VRM (with a Pearson correlation coefficient of **0.519**) is attributed to the HVS included in the ringing region detection model. The added value of the refined ringing annoyance quantification (including the comparison of the variance with the background and the detection of spurious ringing pixels) can be validated by comparing the performance of the metric reported in this paper to its previous version reported in [37]. The gain in performance of the metric reported here over the one reported in [37] corresponds to an increase in the Pearson correlation coefficient from **0.80** to **0.851**. This implies that quantifying ringing annoyance as the absolute intensity variance is effective, but is still too sensitive to remaining texture present in detected ringing regions.

The perceived annoyance level is better addressed by comparing the local variance to the activity of its corresponding local surrounding.

It should be noted that the metric proposed in this paper is only validated for ringing perceived in JPEG compressed images, while ringing is also obviously present in JPEG2000 compressed images. There are, however, a couple of reasons, based on which one can expect a similar performance of our metric on JPEG2000 or H.264 compressed images. First of all, most ringing metrics, and also ours, measure ringing in the spatial domain of the decoded image. As such, these metrics only rely on the characteristics of ringing artifacts (e.g. spatial edge information) rather than on the coding parameters (e.g. DCT coefficients or wavelet coefficients). As a consequence, one would not expect that these metrics need to be intrinsically changed for any of the existing image or video coding standards, but rather can be immediately used or at most need to be slightly modified for measuring ringing artifacts in any type of compressed image. This is confirmed by the claim already made for the metric VRM, namely that it is independent of the particular coding method employed [28], [38]. Additionally, it can be shown that the metric FRRM has a comparable performance for predicting perceived ringing in both JPEG2000 and JPEG compressed images. Indeed, its performance was characterized with a Pearson correlation coefficient of 85% for JPEG2000 compressed images in [14], while we found a Pearson correlation coefficient of 80% for JPEG compressed images in this paper. To illustrate the implementation of our proposed NR ringing metric on a JPEG2000 compressed image, an example is given in Figure 14. It can be seen that the metric successfully identifies and quantifies ringing artifacts in the image. However, to fully evaluate the metric's performance subjective ringing ratings (not the overall quality scores) of JPEG2000 compressed images are needed, which we currently don't have to our availability.

(a) (b)

Fig. 14. Illustration of how the proposed NR ringing metric works on JPEG2000 compression: (a) a JPEG2000 coded image (bit rate 0.099 bpp) from LIVE database [43]; and (b) extracted visible ringing pixels with their visibility indicated with a color ranging from black (no visibility) to yellow for the highest visibility.

Last, but not least, it should be noted that our performance evaluation with a subjective experiment is limited with respect to the amount of test stimuli, the number of human subjects and the display devices used. Adding more experimental data to the performance evaluation would be highly beneficial, but also is very time-consuming. To facilitate further benchmarking of ringing metrics,

apart from developing computational models, future work should also focus on collecting and distributing more reliable subjective data.

4.7 Conclusions

In this paper, a novel no-reference metric for perceived ringing artifacts in compressed images is presented. This metric relies on the existing perceived ringing region detection method [30], [31], and includes ringing annoyance estimation in the perceptually relevant regions in an image. For each individual ringing region, a ringing annoyance score is calculated by first estimating the local visibility of ringing artifacts, and then by comparing it to the local background activity. An overall ringing annoyance score is obtained by averaging the local annoyance scores over all ringing regions. A psychovisual experiment is conducted to measure ringing annoyance subjectively and to validate our proposed ringing metric. The performance of our metric is compared to existing alternatives in literature. It demonstrates that our metric outperforms state-of-the-art metrics in predicting perceived ringing annoyance. Combined with its reliability and computational efficiency, our metric can be a good alternative for real-time implementation.

4.8 References

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

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

[3] 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, April 2004.

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

[5] H. R. Sheikh, A. C. Bovik, and L. K. Cormack, "No-Reference Quality Assessment Using Natural Scene Statistics: JPEG2000," IEEE Transactions on Image Processing, vol. 14, no. 11, pp. 1918-1927, December 2005.

[6] R. V. Babu, S. Suresh and A. Perkis, “No-reference JPEG-image quality assessment using GAP-RBF,” Signal Processing, vol. 87, no.6, pp.1493-1503, 2007.

[7] P. Gastaldo and R. Zunino, “Neural networks for the no-reference assessment of perceived quality,” Journal of. Electronic Imaging, 14 (3), 033004, 2005.

[8] H. R. Wu and M. Yuen, “A generalized block-edge impairment metric for video coding,” IEEE Signal Processing Letters, vol. 4, no. 11, pp. 317–320, 1997.

[9] R. Muijs and I. Kirenko, “A no-reference blocking artifact measure for adaptive video processing,” in Proc. of the 13th European Signal Processing Conference, September 2005.

[10] S. Liu and A. C. Bovik, “Efficient DCT-domain Blind Measurement and Reduction of Blocking Artifacts,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 12, no. 12, pp. 1139-1149, December 2002.
[11] H. Liu and I. Heynderickx, “A Perceptually Relevant No-Reference Blockiness Metric Based on Local Image Characteristics,” EURASIP Journal on Advances in Signal Processing, vol. 2009, 2009.
[12] Z. Wang, H. R. Sheikh and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images", in Proc. IEEE International Conference on Image Processing, vol. 1, pp. 477-480, September 2002.
[13] F. Pan, X. Lin, S. Rahardja, W. Lin, E. Ong, S. Yao, Z. Lu and X. Yang “A locally adaptive algorithm for measuring blocking artifacts in images and videos,” Signal Processing: Image Communication, vol. 19, no. 6, pp. 499–506, 2004.
[14] P. Marziliano, F. Dufax, S. Winkler and T. Ebrahimi, “Perceptual blur and ringing metrics: Application to JPEG2000,” Signal Processing: Image Communication, vol. 19, pp. 163-172, 2004.
[15] H. Tong, M. Li, H. Zhang and C. Zhang, “No-reference Quality Assessment for JPEG2000 Compressed Images,” in Proc. IEEE International Conference on Image Processing, vol. 5, pp. 3539-3542, September 2004.
[16] R. Barland and A. Saadane, “Reference Free Quality Metric for JPEG-2000 Compressed Images,” in Proc. International Symposium on Signal Processing and its Applications, vol. 1, pp. 351-354, August 2005.
[17] X. Li, “Blind image quality assessment,” in Proc. IEEE International Conference on Image Processing, vol. 1, pp. 449-452, Sep. 2002.
[18] M. C. Q. Farias and S. K. Mitra, “No-Reference Video Quality Metric based on Artifact Measurements,” In Proc. IEEE International Conference on Image Processing, vol. 3, pp. 141-144, 2005.
[19] R. Ferzli and L. J. Karam, “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.
[20] K. Zon and W. Ali, “Automated video chain optimization,” IEEE Transactions on Consumer Electronics, vol. 47, pp. 593-603, Aug 2001.
[21] C. C. Koh, S. K. Mitra, J. M. Foley, and I. Heynderickx, “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, March 2005.
[22] M. Shen and C. Kuo, “Review of Postprocessing Techniques for Compression Artifact Removal,” Journal of Visual Communication and Image Processing, vol. 9, no. 1, pp. 2-14, March, 1998.
[23] R. Muijs and J. Tegenbosch, “Quality-Adaptive Sharpness Enhancement Based on a No-Reference Blockiness Metric,” in Proc. Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics, January 2006.
[24] I. O. Kirenko, R. Muijs, and L. Shao, “Coding artifact reduction using non-reference block grid visibility measure,” in Proc. IEEE International Conference on Multimedia and Expo, pp. 469–472, July 2006.

[25] 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
[26] M. Yuen and H. R. Wu, “A survey of hybrid MC/ DPCM/ DCT video coding distortions,” Signal Processing, vol. 70, no. 3, pp. 247-278, November 1998.
[27] D. S. Taubman and M. W. Marcellin, JPEG2000: Image Compression Fundamentals, Standards, and Practice. Norwell, MA: Kluwer, 2001.
[28] S.H. Oguz, Y.H. Hu and T.Q. Nguyen, “Image Coding Ringing Artifact Reduction Using Morphological Post-filtering,” in Proc. IEEE Second Workshop on Multimedia Signal Processing, pp. 628-633, 1998.
[29] X. Feng and J.P. Allebach, “Measurement of Ringing Artifacts in JPEG Images,” in Proc. SPIE, vol. 6076, pp. 74-83, Feb. 2006.
[30] H. Liu, N. Klomp and I. Heynderickx, “Perceptually Relevant Ringing Region Detection Method,” in Proc. of the 16th European Signal Processing Conference, August 2008.
[31] H. Liu, N. Klomp and I. Heynderickx, “A Perceptually Relevant Approach to Ringing Region Detection,” submitted to IEEE Trans. Image Processing.
[32] C. Tomasi and R. Manduchi, “Bilateral filtering for gray and color images,” in Proc. IEEE International Conference on Computer Vision, pp. 836-846, Jan. 1998.
[33] J. Canny, “A Computational Approach to Edge Detection,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 8, no. 6, pp. 679-698, Nov. 1986.
[34] ITU-R Recommendation BT.500-11, Methodology for the subjective assessment of the quality of television pictures, International Telecommunication Union, Geneva, Switzerland, 2002.
[35] R. Franzen: Kodak Lossless True Color Image Suite. Available: http:/ / www.r0k.us/ graphics/ kodak/
[36] A. M. Dijk, J. B. Martens and A. B. Watson, “Quality assessment of coded images using numerical category scaling,” in Proc. SPIE, vol. 2451, pp. 90-101, Mar. 1995.
[37] H. Liu, N. Klomp and I. Heynderickx, "A No-Reference Metric for Perceived Ringing," in Proc. Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics, January 2009.
[38] S. H. Oguz, “Morphological post-filtering of ringing and lost data concealment in generalized lapped orthogonal transform based image and video coding,” Ph.D. dissertation, University of Wisconsin Madison, 1999.
[39] S. Winkler, “Vision Models and Quality Metrics for Image Processing Applications,” Ph.D. dissertation, Dept. Elect., EPFL, Lausanne, 2002.
[40] A. B. Watson, (Ed.), Digital Image and Human Vision. Cambridge MA: The MIT press, 1993.
[41] R. J. Safranek, T. N. Pappas and J. Chen, "Perceptual criteria for image quality evaluation," Handbook of Image and Video Processing, Al Bovik, Editor, Academic Press, 2004.
[42] B. A. Wandell, Foundations of Vision. Massachusetts: Sinauer Associates, Inc., 1995.
[43] 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

Chapter 5

An Efficient Neural Network based No-Reference Approach to an Overall Quality Metric for JPEG and JPEG2000 Compressed Images

Abstract: **Reliably assessing overall quality of JPEG/JPEG2000 coded images without having the original image as a reference is still challenging, mainly due to our limited understanding of how humans combine the various perceived artifacts to an overall quality judgment. A known approach to avoid the explicit simulation of human assessment of overall quality is the use of a neural network. Neural network approaches usually start by selecting active features from a set of generic image characteristics, a process that is to some extent rather ad hoc and computationally extensive. This paper shows that the complexity of the feature selection procedure can be considerably reduced by using dedicated features that describe a given artifact. The adaptive neural network is then used to learn the highly nonlinear relationship between the features describing an artifact and the overall quality rating. Experimental results show that the simplified feature selection procedure in combination with the neural network indeed are able to accurately predict perceived image quality of JPEG/JPEG2000 coded images.**

This chapter is based on the research article submitted as "An Efficient Neural Network based No-Reference Approach to an Overall Quality Metric for JPEG and JPEG2000 Compressed Images" by H. Liu, J. Redi, H. Alers, R. Zunino and I. Heynderickx to Journal of Electronic Imaging.

5.1 Introduction

Understanding and evaluating image quality has become increasingly important for a broad range of applications, such as the optimization of digital imaging systems, the benchmarking of image and video coding algorithms, and the quality monitoring and control in displays [1]. Traditionally, image quality has been evaluated by human subjects, and a mean opinion score (MOS) represents the image quality perceived by the average viewer. When conducted properly, subjective experiments are considered as the most reliable means of assessing image quality. However, performing subjective experiments is very time-consuming, very expensive, and often too slow to be useful in real-world applications. Therefore, during the last decades, a lot of research effort has been devoted to the development of objective metrics that can automatically and quantitatively predict perceived image quality.

Objective metrics reported in literature range from dedicated metrics that measure a specific image distortion to general metrics that assess the perceived overall quality. The various approaches can be classified into full-reference (FR), reduced-reference (RR) and no-reference (NR). FR metrics measure the similarity or fidelity between the distorted image and its original version, where the latter is considered as a distortion-free reference. The most widely used FR metrics are the mean squared error (MSE) and the peak signal-to-noise ratio (PSNR), both aiming at an overall quality assessment [2]. Improved alternatives for these two basic metrics include e.g. the structural similarity (SSIM) index [3] and the visual information fidelity (VIF) index [4]. Since FR metrics require the access to the original, which is mostly not available at the receiver end of an imaging chain, their applicability is limited to in-lab (off-line) testing of image and video processing algorithms. RR metrics are mainly used in scenarios where the reference is not fully available, e.g. in complex communication networks. They make use of certain features extracted from the reference, which are then employed as side information to evaluate the quality of a distorted image (see, e.g., [5]-[7]). Instead in imaging systems with broadcasted content, NR metrics, in which the quality prediction is based on the distorted image only, i.e. without any reference, are more practical. Designing NR metrics, however, is still challenging partly due to the limited understanding of how humans assess image quality.

Recently, considerable progress has been made in the development of NR metrics. Most NR metrics (see, e.g., [8]-[13]) are dedicated metrics measuring a specific type of artifact created by a specific image distortion process. Examples are a metric measuring sensor noise, a metric measuring ringing or blockiness as a consequence of signal compression, or a metric measuring blur generated during acquisition. In such a scenario, the design of the NR metric can make use of the specific characteristics of the artifact, and therefore, generally obtains a higher reliability with respect to the related perceived quality degradation. Specific NR metrics are, for example, used to tune the setting of various parameters in the algorithms of a video chain in current TVs (see, e.g., [14]-[16]). In addition, they can be combined to predict the perceived overall image quality. Various examples of this approach are given in the literature (see, e.g., [17]-[19]); a ringing metric and a blur metric are often combined to assess the overall image quality of wavelet-based compression [18]. This approach, however, largely depends on the reliability of each of the

artifact specific models, and on the efficiency of their combination in a perceptually meaningful way.

An alternative approach for combining individual, dedicated metrics to an estimate for overall image quality is given in [20], in which natural scene statistics are used to blindly determine the overall quality of images compressed by JPEG2000. The approach relies on the assumption that natural images usually exhibit strong statistical regularities, and therefore, reside in a tiny area of the space containing all possible images. Based on this assumption, it quantifies overall image quality by detecting variations in the statistics of image features in the wavelet domain. The approach is promising, but heavily relies on the sophisticated and computationally expensive modeling of natural scene statistics.

Instead of precisely modeling specific artifacts or natural scene statistics, NR image quality assessment has also been formulated as a machine learning problem. This approach has been proved to be effective for the overall quality prediction of a specific distortion type, e.g. JPEG and MPEG-2 compression [21]-[24]. It treats the human visual system (HVS) as a black box, whose input-output relationship between image characteristics and a quality rating is to be learned by computational intelligent tools, such as a neural network (NN). The problem is generally formulated as a regression or function approximation, and the data needed for training are obtained from subjective experiments. During training the error between the desired output (i.e. the subjective quality rating) and the model prediction is minimized. At run time, the properly trained machine implements the resulting model without requiring further computational effort. The critical step in this approach, however, is the selection of the active features, effectively describing the perceived quality. In general, a considerable number of image features is extracted as input to the NN. These so-called common features may be pixel-based as in [21]-[23], or HVS-based as in [24]. In both cases, however, the feature selection requires considerable effort towards optimization, and it is hard to guarantee minimization of the model's complexity at sufficiently high prediction accuracy.

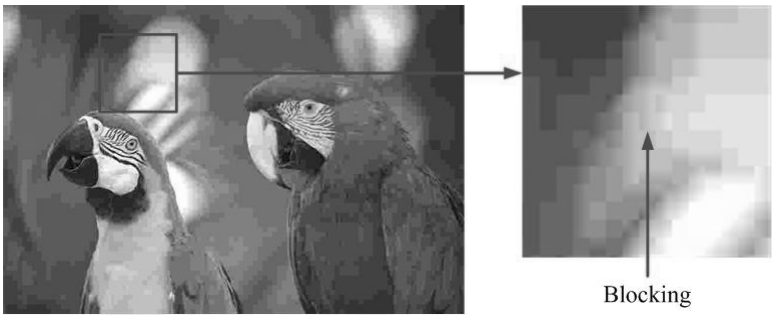
In this paper, we propose to combine the advantages of the two approaches mentioned so far, i.e. the use of (aspects of) artifact-specific metrics as features, and the use of a NN to assess the overall perceived quality. In practice, the approach has two components: first, we calculate the feature(s) describing the most relevant artifact in JPEG/ JPEG2000 compressed images, and second, we use an adaptive NN to learn the highly nonlinear relationship between the feature(s) and the overall quality rating. The use of features dedicated to a single artifact is motivated from results in literature reporting a high correlation between a specific artifact type and the overall quality of JPEG/ JPEG2000 compressed images [25]-[27]. Our novel approach is highly efficient for two reasons. First, it calculates features based on artifact characteristics, and so, this avoids a lengthy and tedious feature selection procedure. In addition, the usefulness of the selected features for predicting image quality is already known from literature. Second, it leaves the simulation of the HVS for the perceived overall image quality to the NN, and as such this part of the model is reduced after training to the implementation of a simple algorithm at run-time. It should be noted that the whole process only uses the luminance component of the images, as such further reducing the computational load.

Section II of this paper discusses the feature-extraction process, and derives the numerical descriptors for the NN that are based on simple yet efficient metrics for

both blockiness and blur artifacts. Section III describes the actual quality prediction tool, which relies on empirical training of a neural network. Section IV presents the overall performance of the proposed NR JPEG and JPEG2000 metrics and a comparison with metrics existing in literature. Section V is devoted to a discussion of the specific added value of the proposed approach.

5.2 Feature Extraction and Description

The literature shows that the overall quality of JPEG compressed images is highly correlated with the occurrence of blocking artifacts [25], while the overall quality of JPEG2000 compressed images is highly correlated with the occurrence of blur [18]. A blocking artifact manifests itself as an artificial discontinuity in the image content, which is a direct consequence of the fact that the quantization in JPEG is block-based and that the blocks are quantized independently. A blur artifact occurs in JPEG2000 compressed images mainly due to the loss of high frequency transform coefficients in the wavelet-based coding, as a result of which the image signal is smoothened. Figure 1 illustrates the occurrence of blocking artifacts in a JPEG compressed image, and of blur artifacts in a JPEG2000 compressed image, respectively.



(a)

(b)

Fig. 1. Illustration of the occurrence of compression artifacts: (a) blocking artifacts in an image JPEG compressed with a bit rate of 0.21 bits per pixel (bpp), and (b) blur artifacts in an image JPEG2000 compressed with a bit rate of 0.099 bpp.

Quality degradation as a consequence of compression should then easily be predictable from the extraction of blockiness/ blur related image features, provided that an adaptive NN is used to empirically learn the highly nonlinear relationship between these artifact-oriented features and the overall quality rating. To efficiently characterize the local behavior of artifacts and thus to feed the neural network with relevant features for image quality prediction, a gradient-based feature extraction scheme is proposed. It contains three basic components: (1) the localization of the artifacts, (2) the local feature extraction using local gradients in relation to their neighborhood, and (3) the assembling of a global statistical descriptor as input to the neural network. The implementation of each of these steps is detailed below.

5.2.1 Local Feature Extraction: JPEG

Due to the underlying coding algorithm for JPEG compression, the spatial location of blocking artifacts is very regular. In principle, they occur on a grid of blocks of 8×8 pixels, starting at the top-left corner of an image. In common applications, however, grid sizes may differ and starting positions may shift, either due to perturbations in the incoming signal or as a consequence of spatial scaling. In such a scenario, a (naïve) NR metric might run the risk of calculating blockiness at wrong pixel positions, and therefore might incur in a dramatic degradation in accuracy [13]. To ensure that the metric is calculated exactly at block boundaries, a grid detector can be adopted. The research presented in this paper implements the blocking grid detection method proposed in [13]. It is, however, worth stressing that the feature-extraction approach is independent of the particular choice of grid detector, hence any alternative approach (e.g. the one described in [16]) can be applied. The blocking grid detector first maps an image onto a 1-D signal profile, in which the periodic property of blocking artifacts is maintained. Then the exact block size as well as the grid offset is easily extracted from the discrete Fourier transform (DFT) of this 1-D signal profile.

When the blocking artifacts are (exactly) located, the related feature can be extracted. In this paper, the feature for the JPEG compressed images is based on the visual strength of a blocking artifact within a local area of the image content [13]. Since a blocking artifact is a local edge that stands out from its spatial vicinity, it can be simply defined relating the energy present in the gradient at the artifact to the energy present in the gradient in its neighboring pixels. When the luminance channel of an image of M×N (height × width) pixels is denoted as I(i, j) for i∈[1, M], j∈[1, N], the local blockiness $L_{\text{blockiness-h}}$ along the horizontal direction at location (i, j) is quantified as

$$L_{\text{blockiness-h}}(i, j) = \frac{G_h(i, j)}{\frac{1}{2n}\sum_{x=-n,\ldots,n,x\neq 0} G_h(i, j+x)} \qquad (i, j) \in \{\text{blocking grid}\} \qquad (1)$$

where $G_h(i, j)$ indicates the gradient map along the horizontal direction, and is computed as

$$G_h(i, j) = \left|I(i, j+1) - I(i, j)\right|, j \in [1, N-1] \qquad (2)$$

where n determines the size of the template used to describe the local content. The size is determined as a balance between sufficient information of the local content, while avoiding noise from content too far away. In our experiments we used n=3, being equal to half the amount of pixels between two blocking edges.

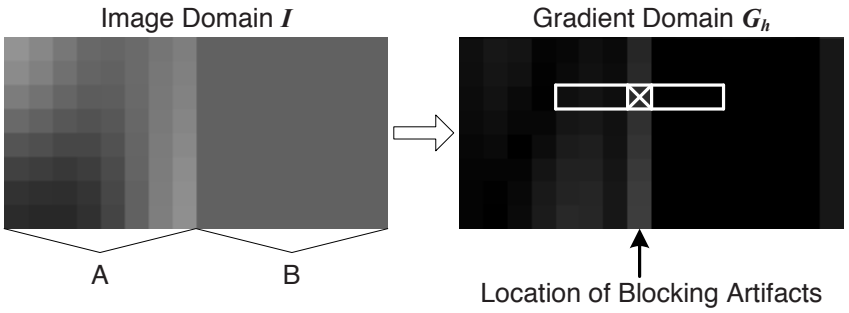


Fig. 2. Illustration of the template for calculating the local blockiness: (a) two adjacent 8×8 blocks (i.e. A and B) extracted from a real JPEG image, and (b) the gradient profile of the image patch of (a).

An example of the template for calculating $L_{blockiness\text{-}h}$ is shown in Figure 2, where two adjacent blocks of 8×8 pixels (i.e. A and B) are extracted from a real JPEG image. The local blockiness along the vertical direction $L_{blockiness\text{-}v}$ can be calculated similarly. The higher the values of $L_{blockiness\text{-}h}$ and $L_{blockiness\text{-}v}$ are, the larger the distortion of the blocking artifact is. It should, however, be noted that this does not necessarily mean that the blocking artifact is also more visible. The local visibility of a blocking artifact may be affected by texture and luminance masking, which typically occur in the HVS. It has been shown in literature that taking into account these masking effects can be greatly beneficial for the prediction performance of a dedicated NR blockiness metric [13]. However, modeling the HVS introduces more computational power. So, in this paper we avoid the calculation of masking, and rely on the NN to learn the unknown functional relationship between the extracted gradient-based features and the rating of overall image quality.

5.2.2 Local Feature Extraction: JPEG2000

In JPEG2000 compression, blur artifacts are perceptually prominent along edges or in textured areas. Hence, in this paper, the local feature extracted for the JPEG2000 compressed images calculates the degree of blur at an edge within a local area of image content. Literature offers a wide variety of techniques to detect strong edges, and consequently to identify the spatial location of blur artifacts (e.g. [28] and its references). The implemented approach uses a straightforward Sobel edge detector resulting in a gradient image. The location of strong edges is then extracted by applying a threshold to this gradient image (as such removing noise and insignificant edges). The threshold value is automatically set depending on the image content (e.g. using the mean of the gradient magnitude squared image).

We then use a novel, simple, yet efficient measure for the blur of all detected edges. Instead of calculating the distance between the start and end position of an edge (as proposed in [8]), edge blur is locally defined in the gradient domain as the sharpness of the edge related to its surrounding content within a limited extent. When describing blur simply as the relative gradient energy of an edge compared to its direct vicinity, it can be quantified in the same manner as used in (1), i.e.:

$$L_{blur\text{-}h}(i, j) = \frac{G_h(i, j)}{\frac{1}{2n} \sum_{x=-n,\ldots,n, x\neq 0} G_h(i, j+x)} \qquad (i, j) \in \{\text{strong edges}\} \qquad (3)$$

where $L_{blur\text{-}h}$ indicates the local blur along the horizontal direction and n, representing the size of the template, has the same value as in equation (1). $L_{blur\text{-}v}$, i.e. the local blur in the vertical direction, can be calculated similarly. The lower the value of $L_{blur\text{-}h}$ and $L_{blur\text{-}v}$, the larger the distortion of the blur artifact is. Figure 3 explains the reasoning behind the proposed approach of using gradient energy to detect blur. Figures 3(a) and (c) show a detected edge (i.e. at location (113, 259)) in the JPEG2000 compressed image of Figure 1(b), and its intensity profile over the pixels in its direct vicinity, respectively. Figure 3(b) and (d) show the corresponding edge in the original uncompressed image of Figure 1(b), and its intensity profile over the pixels in its direct vicinity. The difference in sharpness between the two edges is clearly revealed in the gradient domain (see Figures 3(e) and (f)). In correspondence, the values of $L_{blur\text{-}h}$ indicate that the edge of Figure 3(a) is more blurred than the edge of Figure 3(b).

5.2.3 Global Descriptor of the Image Features

Once the local features related to blocking/ blur artifacts are explicitly extracted and calculated for each JPEG/ JPEG2000 compressed image, the results can be visualized in a spatially varying feature map. An example is given in Figure 4 for a JPEG and JPEG2000 compressed image, respectively. Figures 4(b) and (d) illustrate the location of the artifacts in the horizontal direction, and the intensity at each pixel indicates the local degree of distortion; i.e. the higher the intensity, the larger the distortion is. The location and intensity of the artifacts in vertical direction can be obtained in a similar way.

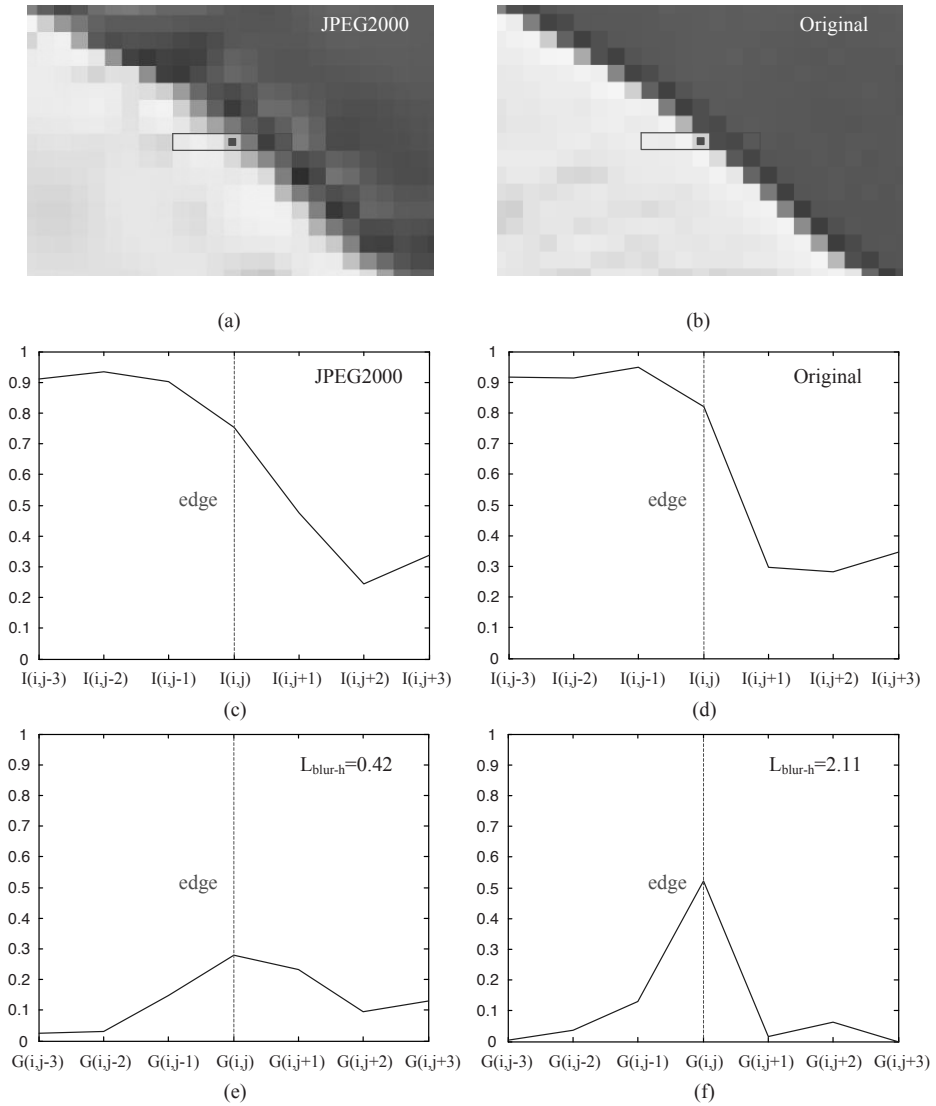


Fig. 3. Illustration of the calculation of local blur: (a) image patch extracted from the JPEG2000 compressed image of Figure 1(b) [the red dot indicates the location of the detected edge at (113, 259) in Figure 1(b), and the template indicates the area in which the local blur is calculated for this edge], (b) the image patch of the original uncompressed image corresponding to (a), (c) the intensity profile over the pixels within the template of (a), (d) the intensity profile over the pixels within the template of (b), (e) the gradient profile of (c), and (f) the gradient profile of (d).

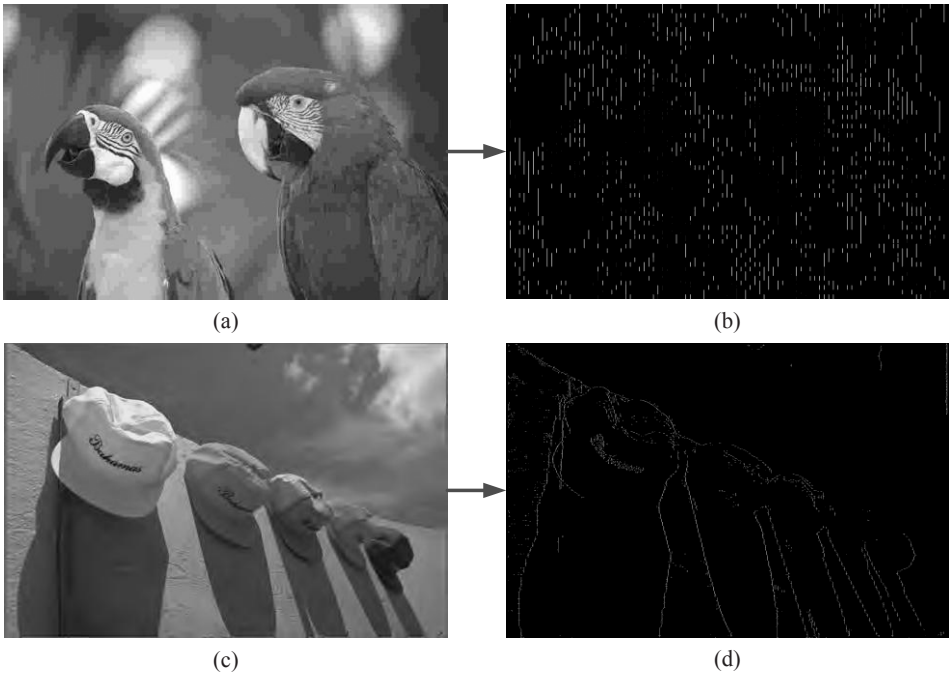


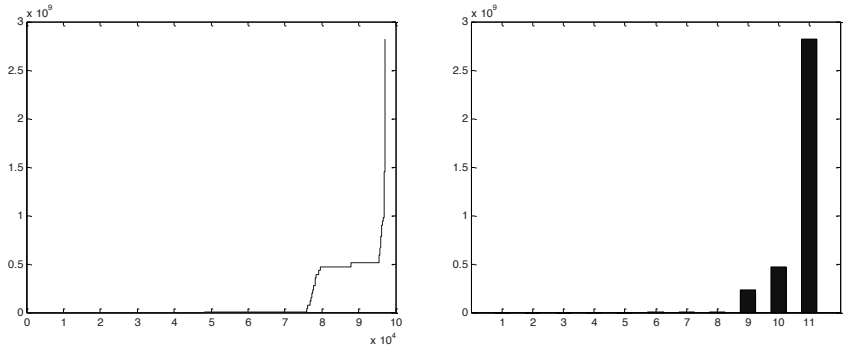
Fig. 4. Visualization of the extracted features: (a) the JPEG compressed image of Figure 1(a), (b) the feature map of (a) showing the location and the degree of distortion of the blocking artifacts in the horizontal direction, (c) the JPEG2000 compressed image of Figure 1(b), and (d) the feature map of (c) showing the location and the degree of distortion of the blur artifacts in the horizontal direction. In (b) and (d), the higher the intensity, the larger the distortion is.

Direct application of all extracted feature values as input to a NN is problematic, since the dimension of the space of these values is often too large, and as such inappropriate for the network in terms of training. Existing approaches to reduce the number of feature values (see, e.g. [21]-[24]) usually calculate a statistical descriptor that characterizes the whole image. This descriptor is a single vector, which needs to be associated with the single quality score generated by human subjects. In this paper, the statistical description of an image feature as proposed in [21]-[23] is adopted. It unifies the local feature values of an image to a single vector using percentiles. Having computed the feature values f_i (i=1, ..., N_F) per image (i.e. $L_{blockiness}$ calculated in both the horizontal and vertical direction on the blocking grid or L_{blur} calculated in both the horizontal and vertical direction on the detected edges), these values are sorted in ascending order of magnitude. The envelope of the obtained distribution is then expressed in a global descriptor **f** by taking 11 of its percentiles φ :

$$\mathbf{f}=\{\varphi_\alpha;\alpha\in\{0,10,20,30,40,50,60,70,80,90,100\}\};\quad \varphi_\alpha=\left\lfloor\frac{N_F}{100}\alpha+\frac{1}{2}\right\rfloor \tag{4}$$

Figure 5 illustrates the formation of the global descriptor of an image feature. Compared to simply taking the average of the feature values, this spatial pooling

strategy allows feeding the nonlinear regression with a more complete overview of the amount and behavior of the considered distortion in the image.



(a) (b)

Fig. 5. Illustration of the formation of the global descriptor of an image feature: (a) the feature values (i.e. the blocking artifacts along the horizontal and vertical direction) extracted from the JPEG compressed image in Figure 1(a) and sorted in ascending order of magnitude, and (b) the global descriptor of the image feature, taking 11 percentiles of the distribution of (a).

5.3 NR Image Quality Estimator Based on A Neural Network

As reported in literature already (see, e.g. [21], [22], [23], [24], [29]) we implement a NN to approximate the functional relationship between the image features and the related quality score. The main difference with earlier contributions to literature is that in our case the input feature vector to the NN contains descriptors of the actually occurring artifacts (see Section II). The NN aims to mimic the mechanism of quality perception and avoids an explicit model of the HVS, thus reducing the number of assumptions typically required to model perceived quality analytically. In this paper, a feed-forward NN is employed to operate on the feature vector extracted from JPEG/ JPEG2000 images. The implementation of this NN is already described in more detail in [21]-[23], and is only briefly repeated here.

A feed-forward NN aims at implementing a stimulus-response behavior by arranging several elementary units ("neurons") into a layered structure, which does not allow any feedback between layers. Each neuron involves a simple, nonlinear transformation of weighted inputs, and the nonlinearity is often performed by a sigmoidal function. The multilayer Perceptron (MLP) paradigm [30] belongs to this type of networks, and it has been proved to be able to perform effectively in scenarios where the target mapping function can be determined by a few computing units with global scope. It intrinsically implements a series expansion of n_h basis functions a_h (i.e. sigmoids), which can be generally expressed as:

$$y(\mathbf{x}) = w_0 + \sum_{h=1}^{n_h} w_h a_h(\mathbf{x}) \tag{5}$$

where **x** indicates the stimulus vector with its output value $y(\mathbf{x})$, and the coefficients of **w** are called the network "weights" and need to be adjusted during the training phase. The basic scheme of (5) is usually enhanced by applying a sigmoidal nonlinearity to the output value.

The circular back-propagation (CBP) network [31] improves the conventional MLP paradigm by adding one more input value, which is the sum of the squared values of all the network inputs. For an input stimulus vector $\mathbf{x}=\{x_1, \ldots, x_{ni}\}$, the input layer connects the n_i values to each neuron of the "hidden" layer. The j-th "hidden" neuron performs a nonlinear transformation of a weighted combination of the input values with coefficients ("weights") $\mathbf{w}_{j,i}$ (j=1, …, n_h, and i=1, …, n_i):

$$a_j = sigm(w_{j,0} + \sum_{i=1}^{n_i} w_{j,i} \cdot x_i + w_{j,n_i+1} \sum_{i=1}^{n_i} x_i^{\,2}) \tag{6}$$

where $sigm(z)=(1+e^{-z})^{-1}$, $\mathbf{w}_{j,0}$ is a bias term, and a_j is the neuron activation (i.e. the output of the basis function). The output layer provides the final network response, y_k, (k = 1 in the case of image quality assessment):

$$y_k = sigm(w_{k,0} + \sum_{j=1}^{n_h} w_{k,j} \cdot a_j) \tag{7}$$

where $\mathbf{w}_{k,j}$ and $\mathbf{w}_{k,0}$ represent the output coefficients and the output bias, respectively.

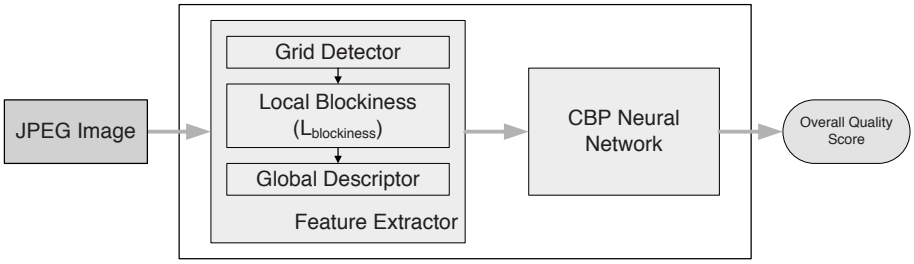
The resulting CBP network can map both linear and circular separation boundaries [31]. The additional input value enhances the overall representation ability of the network, while not affecting the properties of the MLP structure (e.g. $\mathbf{w}_{j,ni+1}$=0 reduces a CBP network to a classical MLP). Since the actual coefficients of $\mathbf{w}_{j,ni+1}$ are determined by the empirical training process, the selection between a conventional MLP and a CBP model is entirely data-driven and does not require any a priori assumption. Such an adaptive behavior makes CBP networks appropriate for perception related problems, whose underlying structure is often obscure.

The degrees of freedom of the NN that need to be fitted are the depth n_h of the series expansion and the weighting coefficients within each neuron. To determine the former quantity, literature provides both theoretical [32] and practical [33] criteria to ensure prediction accuracy, while minimizing the risk of over-fitting training data. In this paper, we follow an empirical approach [33] mainly due to its simplicity and proved effectiveness. Once the number of network neurons (i.e. n_h) is decided, a fitting process tunes the set of weights in such a way that the network optimizes the desired input-output mapping, minimizing a cost function which implements the mean square error between the predictions and the subjective quality scores.

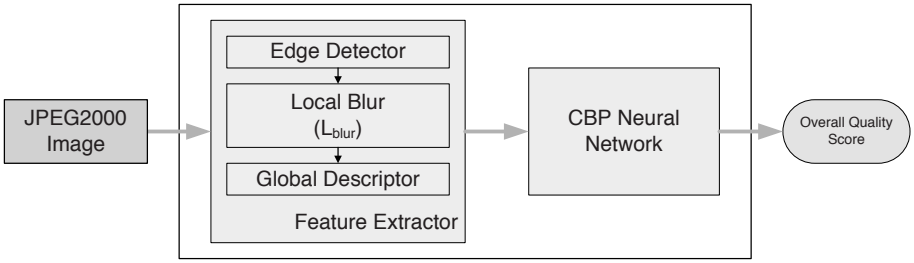
5.4 Evaluation of the Overall Metric Performance

5.4.1 Test Environment

Figure 6 illustrates the schematic overview of the proposed NR metric for the perceived overall quality assessment of JPEG/ JPEG2000 compressed images. It should be mentioned that the approach discussed in this paper still treats the JPEG and JPEG2000 images separately, since a NR metric is feasible only when the prior knowledge about the image distortion process is available [1]. But, we envision that by including a classification algorithm of JPEG and JPEG2000 (see e.g. in [42]) the system can automatically select the appropriate metric to use on any compressed image. This is, however, outside the scope of this paper. In our experiments, for each image, a vector containing eleven percentiles of the distribution of the local blockiness/ blur features was calculated as the input to the NN. The CBP network was equipped with three hidden neurons and trained with the back-propagation [34] algorithm.



(a)



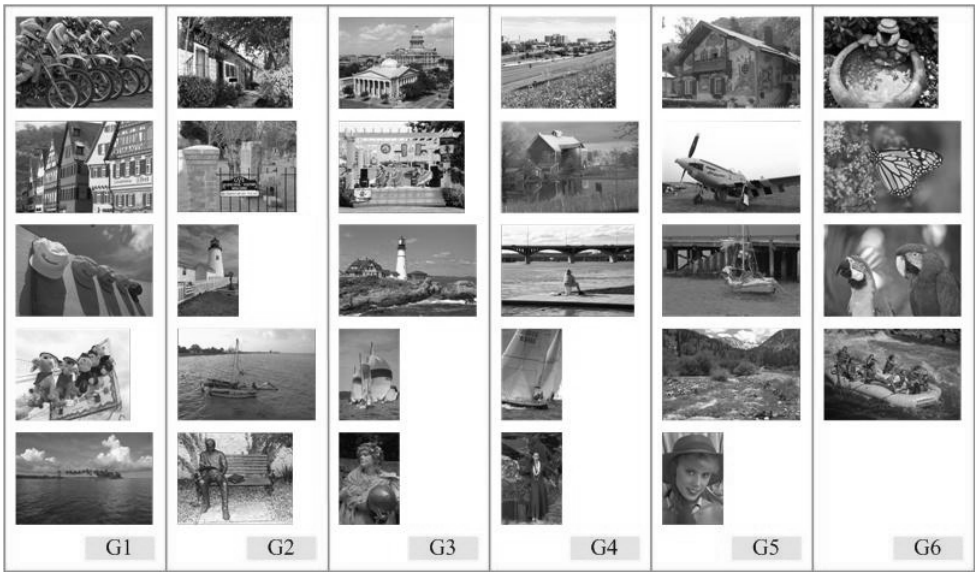
(b)

Fig. 6. Schematic overview of the proposed NR metric for the perceived overall quality assessment of JPEG/ JPEG2000 compressed images: (a) the NR JPEG metric, and (b) the NR JPEG2000 metric.

To evaluate the performance of the proposed approach, the LIVE image quality assessment database [35] was used. It consists of a set of twenty nine high-resolution and high-quality color source images that reflect adequate diversity in image content. These images were compressed using JPEG at a bit rate ranging from 0.15 bits per pixel (bpp) to 3.34 bpp, resulting in a database of 233 JPEG

compressed stimuli (including the originals). The same source images were also compressed using JPEG2000 at a bit rate ranging from 0.028 bpp to 3.15 bpp, yielding a database of 227 JPEG2000 compressed stimuli (including the originals). An extensive psychovisual experiment was conducted to assign a difference mean opinion score (DMOS) to each stimulus. The DMOS was measured on a continuous linear scale that was divided into five intervals marked with the adjectives "Bad", "Poor", "Fair", "Good" and "Excellent".

So far, empirically measuring the error on the test data (e.g. by cross-validation [36]-[38]) was proved to be the most reliable method to achieve an accurate approximation of the performance of a NN system. In our performance evaluation, a K-fold cross-validation method [38] was adopted (see also our previous experimental setup reported in [41]). It randomized the statistical design problem by repeatedly splitting the available data in a training set and a test set. Figure 7 illustrates the experimental setup, in which the source images were divided into six groups. The entire procedure included six different trials, and for each trial (hereafter referred to as "run") five groups of source images were used for training and the remaining one group of source images was used for testing. It is noteworthy that none of the stimuli used for testing ever entered any step of the training process. This way of working served to assess the generalization of the system performance empirically. It resulted in (an average of) 194 stimuli for training and (an average of) 39 stimuli for testing in the JPEG database for evaluating our JPEG metric. Additionally, we had (an average of) 189 stimuli for training and (an average of) 38 stimuli for testing in the JPEG2000 database for evaluating our JPEG2000 metric.



(a)

	Training	Test
Run 1	G1 G2 G3 G4 G5	G6
Run 2	G1 G2 G3 G4 G6	G5
Run 3	G1 G2 G3 G5 G6	G4
Run 4	G1 G2 G4 G5 G6	G3
Run 5	G1 G3 G4 G5 G6	G2
Run 6	G2 G3 G4 G5 G6	G1

(b)

Fig. 7. Experimental setup for the K-fold cross-validation: (a) source images of the LIVE database [35] assigned over six groups, and (b) distribution of the source images assigned over the training and test set for each of the six runs.

5.4.2 Overall Metric Performance

As prescribed by the video quality experts group (VQEG) [39], the performance of our approach was evaluated with respect to its ability to predict subjective quality ratings (the DMOS). Two statistical tools usually employed in literature were adopted to characterize the prediction ability: i.e. the Pearson linear correlation coefficient, and the root mean square error (RMSE). The corresponding correlation coefficients and RMSE are listed in Table 1 and 2 for the JPEG and JPEG2000 compressed images, respectively. In both cases, our proposed metric consistently resulted in a high prediction performance over all (six) runs. The NR JPEG metric yielded an averaged Pearson correlation coefficient of **0.9623** (with a highest value of 0.975 and a lowest value of 0.953), and an averaged RMSE of **0.109** on a normalized scale [0, 1] (with a highest value of 0.127 and a lowest value of 0.084). The NR JPEG2000 metric provided an averaged Pearson correlation coefficient of **0.930** (with a highest value of 0.942 and a lowest value of 0.923), and an averaged RMSE of **0.139** on a normalized scale [0, 1] (with a highest value of 0.155 and a lowest value of 0.115).

Table 1. Performance of the proposed NR JPEG metric per run and averaged over all runs in terms of Pearson correlation coefficient and RSME.

NR JPEG Metric	Pearson Correlation Coefficient	Root Mean Square Error (RMSE) (normalized score scale [0, 1])
Run 1	0.963	0.119
Run 2	0.956	0.127
Run 3	0.962	0.115
Run 4	0.953	0.116
Run 5	0.975	0.084
Run 6	0.965	0.096
MEAN	**0.962**	**0.109**

Table 2. Performance of the proposed NR JPEG2000 metric per run and averaged over all runs in terms of Pearson correlation coefficient and RSME.

NR JPEG2000 Metric	Pearson Correlation Coefficient	Root Mean Square Error (RMSE) (normalized score scale [0, 1])
Run 1	0.942	0.115
Run 2	0.934	0.138
Run 3	0.926	0.144
Run 4	0.924	0.155
Run 5	0.923	0.145
Run 6	0.925	0.139
MEAN	**0.930**	**0.139**

5.4.3 Comparison to Alternative Metrics

In the image quality community, researchers are accustomed to compare their metrics to alternatives available in the literature. It is, however, important to note that the performance of these metrics needs to be evaluated in a comparative setting, so that their strengths and weaknesses are fairly analyzed. In this respect, apart from only listing the numerical results (e.g. Pearson correlation coefficient) of the metrics in comparison, we also address some important issues behind these values.

An objective metric is conventionally validated through quantifying the correlation between its predicted values and the subjective quality scores. This correlation, however, can be calculated under three different testing conditions (TC) as reported in literature. The first one is to directly calculate the linear correlation between the metric's predictions and the subjective data (see e.g. [11], [13], [40] and hereafter referred to as TC1). This method is often used in metric comparison to better visualize differences in performance. The second method is suggested by the VQEG [39], and applies a nonlinear mapping function (e.g. a logistic function) to fit the metric's results to the DMOS before computing the correlation (referred to as TC2). A sophisticated nonlinear fitting (often including various parameters) accounts for a possible saturation effect, and usually yields higher correlation coefficients in absolute terms. However, in terms of performance comparison, it generally keeps the relative differences between the metrics as computed under TC1 [11]. A more demanding testing condition is cross-validation, in which the dataset is partitioned into complementary subsets: one for model calibration and the other for validation (see e.g. [23], [24], [29], and hereafter referred to as TC3). Therefore, it should be noted that simply comparing the reported correlation coefficients of different metrics is not meaningful, unless they are evaluated with the same database under the same testing condition. Even under the same testing condition, the quantitative comparison between metrics may be biased due to e.g. a different selection of the disjoint sets for training and testing (e.g. under TC3).

Here, we compare the proposed JPEG and JPEG2000 metrics to state-of-the-art NR metrics in terms of performance. Issues related to computational complexity will be discussed in Section V.A. For practical reasons the NR metrics used for comparison are limited to four JPEG metrics and three JPEG2000 metrics. To further

compare the performance of our NR metric with respect to RR and FR metrics, we also include two RR metrics and three FR metrics well-known in literature. However, is should be noted that a direct comparison of NR metrics to RR/ FR metrics is not completely fair, since the design of reliable NR metric is more challenging, and often a NR metric is the only available option in real-time applications. Tables 3 and 4 list the Pearson correlation coefficient and the corresponding testing environment for these metrics. It can be seen that the prediction performance of our proposed JPEG and JPEG2000 metrics is slightly better than currently leading NR metrics. For the JPEG metric, our proposed approach even outperforms some of the existing RR and FR metrics, and comes close in performance to the best in class of these metrics. For the JPEG2000 metric, our approach is slightly underperforms with respect to RR and FR metrics currently known in literature.

Table 3. Performance of state-of-the-art NR JPEG metrics.

JPEG Metric		**Pearson Correlation Coefficient**	**Testing Environment**
NR	Liu et al [13]	0.918	LIVE JPEG – TC1
	Wang et al [17]	0.931	LIVE JPEG – TC2
	Gastaldo et al [23]	0.943	LIVE JPEG – TC3
	Babu et al [24]	0.932	LIVE JPEG – TC3
	Proposed	**0.962**	LIVE JPEG – TC3
RR	Li et al [5]	0.889	LIVE JPEG – TC2
	Carnec et al [6]	0.972	LIVE JPEG – TC2
FR	PSNR [2]	0.901	LIVE JPEG – TC2
	SSIM [3]	0.979	LIVE JPEG – TC2
	VIF [4]	0.980	LIVE JPEG – TC2

Table 4. Performance of state-of-the-art NR JPEG2000 metrics.

NR JPEG2000 Metric		**Pearson Correlation Coefficient**	**Testing Environment**
NR	Marziliano et al [18]	0.850	LIVE JP2K – TC3
	Sheikh et al [20]	0.910	LIVE JP2K – TC3
	Sazzad et al [19]	0.930	LIVE JP2K – TC3
	Proposed	**0.930**	LIVE JP2K – TC3
RR	Li et al [5]	0.957	LIVE JP2K – TC2
	Carnec et al [6]	0.957	LIVE JP2K – TC2
FR	PSNR [2]	0.904	LIVE JP2K – TC2
	SSIM [3]	0.971	LIVE JP2K – TC2
	VIF [4]	0.979	LIVE JP2K – TC2

5.5 Evaluation of Specific Metric Components

5.5.1 Reduction in Computational Complexity

From a practical point of view, it is highly desirable to develop a NR metric that is easy to implement, computationally efficient, and uses only a few parameters. In this section, we specifically elaborate on the reduction in computational complexity of the proposed approach compared to existing alternatives in the literature.

Many NR metrics for JPEG compression are reported in literature, and the most successful ones are listed in Table 3. Our proposed metric, however, outperforms these alternatives in terms of simplicity, still obtaining a high reliability, as shown in Table 3. The metric of [13] explicitly models the HVS via texture and luminance masking. To keep the computational load of the metric within reasonable limits, both masking processes were heavily simplified, and that limited the metric's performance. Compared to the metrics of [17], [23], and [24], which involved an extensive feature computation or selection stage (e.g. in [17] both blockiness and signal activities were calculated, in [23] a large number of general pixel-based features were extracted, and in [24] a variety of HVS related features were computed), our metric clearly shows its advantage by only simply calculating the local blockiness in an computationally efficient way.

Progress in NR metrics for JPEG2000 compression is limited, mainly because the various artifacts are inherently content-dependent, and so, difficult to be detected and modeled. Researchers have taken different approaches to this problem. The metric of [18] is a well-known metric that attempts to predict the overall quality of JPEG2000 compressed images by modeling their most relevant artifact, which is blur. It is a very simple metric, however, its reported performance is limited, since the averaged blur is mapped to the quality scores with only a simple nonlinear transformation. Our metric clearly outperforms the metric of [18] (see Table 3), yet without introducing additional computational cost. The metric of [20] adopts natural scene statistics for measuring the image quality, which, however, often requires sophisticated modeling to achieve a reliable metric. A recently proposed NR metric for JPEG2000 compression [19] reports a high correlation with the LIVE database [35], but contains an intensive feature extraction stage (i.e. eight different spatial features) and a complex parameter optimization procedure (i.e. nine model parameters) to combine these features. Thus, our metric outperforms the alternatives of [19] and [20] in implementation complexity and computational efficiency.

5.5.2 The Added Value of Using a Neural Network

The promising performance of the NR metrics, proposed in this paper, is primarily achieved by the combination of two essential components: (1) a simplified feature extraction that largely reduces the computational complexity and avoids multiple-feature modeling, and (2) a powerful NN to map the extracted feature to a quality rating. To validate the added value of including a NN, additional experiments were conducted, in which the neural network was omitted. Instead, the averaged feature value (i.e. the mean of the calculated local blockiness for the JPEG metric, and the

mean of the calculated local blur for the JPEG2000 metric) was used as the metric's output. To make a fair comparison to our original NN based metric, we evaluated the resulting metrics under three different testing conditions, as mentioned in Section IV.C. Tables 5 and 6 list the correlation coefficient and the RMSE for each of the testing conditions. It should be mentioned that the logistic function suggested by VQEG [39] is conventionally used in both TC2 and TC3. The nonlinear regression function transforms the metric's predictions to a set of predicted MOS values (i.e. $DMOS_p$), which are then compared to the actual DMOS values. The three-parameter logistic function is expressed as:

$$DMOS_p = \frac{b1}{1+\exp(-b2\times(Metric-b3))} \tag{8}$$

Table 5. Performance when using only the averaged blockiness feature value for the three testing conditions described in Section IV.C.

Testing Environment		Pearson Correlation Coefficient	Root Mean Square Error (RMSE)
LIVE JPEG - TC1		**0.661**	0.349
LIVE JPEG - TC2		**0.906**	0.247
LIVE JPEG – TC3	**RUN1**	0.875	0.250
	RUN2	0.929	0.188
	RUN3	0.912	0.245
	RUN4	0.926	0.237
	RUN5	0.872	0.238
	RUN6	0.907	0.211
	MEAN	**0.904**	0.228

Table 6. Performance when using only the averaged blur feature value for the three testing conditions described in Section IV.C.

Testing Environment		Pearson Correlation Coefficient	Root Mean Square Error (RMSE)
LIVE JPEG2000 - TC1		**0.715**	0.257
LIVE JPEG2000 - TC2		**0.741**	0.221
LIVE JPEG2000 – TC3	**RUN1**	0.811	0.198
	RUN2	0.779	0.230
	RUN3	0.769	0.214
	RUN4	0.818	0.213
	RUN5	0.669	0.259
	RUN6	0.675	0.255
	MEAN	**0.754**	0.228

The experimental results show that the simple, single feature without the use of a NN hardly achieves a reliable metric, even not after a complex fitting of the metric's output values to the corresponding subjective ratings. Of course, the performance of these simple metrics can be improved by adding properties of the HVS and by explicitly modeling the inherent artifacts (see e.g. [9], [12], [13]). This generally

yields metrics that can be used to assess the overall image quality, but at the expense of complex HVS modeling and an extensive parameter optimization process. For real-world applications, our proposed approach including a NN tends to be an efficient and inexpensive solution.

5.5.3 Limitations and Future Research

The NR metrics, proposed in this paper, aim for assessing the perceived *overall* quality of JPEG and JPEG2000 compressed images. To achieve a simple yet efficient metric, especially for real-time processing, we kind of neglect the occurrence of and interaction between various artifacts that may occur simultaneously in an image, thus affecting the perceived overall quality. In our case, only the most relevant artifact is extracted to predict the overall image quality, and we fully rely on the NN to approximate the unknown relationship between this single feature and the quality rating. As a consequence, the proposed NR metrics intrinsically exhibit two major drawbacks: (1) the local distortion values (simply calculated) are not necessarily in agreement to what the human eye perceives, and thus cannot be used to precisely reflect the local annoyance of a perceived artifact, and (2) the perceived annoyance of other artifacts, e.g. ringing in JPEG and JPEG2000 compression, cannot be assessed by the overall metric. Being able to quantify the annoyance of more types of artifacts is of fundamental importance to e.g. noise reduction in image/ video enhancement. It requires research in the design of dedicated metrics (see e.g. [8]-[13]) to detect and estimate the local annoyance of a specific artifact type. However, in current visual communication systems, predicting the perceived overall quality in real-time without compromising the system's complexity is very valuable, but still challenging. We feel that designing an NR metric based on a NN is promising in terms of predicting the overall image quality. We are continuing our efforts into designing NR metrics for more types of distorted images, such as blur, noise and wireless channel errors.

5.6 Conclusions

In this paper, we provide an efficient NR approach for the perceived overall quality assessment of JPEG/ JPEG2000 compressed images. Its reliable prediction ability at a largely reduced computational cost is achieved by skillfully combining a simplified feature extraction strategy with an adaptive neural network. The first component efficiently selects and calculates the most relevant feature representative for the overall image quality, and thus avoids explicitly modeling the occurrence of and interaction between various artifacts inherent in a distorted image. The latter component, subsequently, is used to empirically learn the highly nonlinear relationship between the relevant feature and the overall image quality rating. The resulting NR JPEG and JPEG2000 metrics are validated with subjective data under a critical cross-validation condition, and are fairly compared to several alternative metrics existing in literature.

5.7 References

[1] Z. Wang and A. C. Bovik, Modern Image Quality Assessment. Synthesis Lectures on Image, Video, & Multimedia Processing, Morgan & Claypool, San Refael, Calif, USA, 2006.

[2] Z. Wang and A. C. Bovik, “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, Jan. 2009.

[3] 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, April 2004.

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

[5] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation,” IEEE Journal of Selected Topics in Signal Processing, Special issue on Visual Media Quality Assessment, vol. 3, no. 2, pp. 202-211, Apr. 2009.

[6] M. Carnec, P. Le Calleta and D. Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference,” Signal Processing: Image Communication, vol. 23, no. 4, pp. 239–256, Apr. 2008.

[7] U. Engelke, M. Kusuma , H. Zepernick a, M. Caldera, “Reduced-reference metric design for objective perceptual quality assessment in wireless imaging,” vol. 24, no. 7, pp. 525-547, Signal Processing: Image Communication, 2009.

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

[9] R. Ferzli and L. J. Karam, “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] X. Feng and J.P. Allebach, “Measurement of Ringing Artifacts in JPEG Images,” in Proc. SPIE, vol. 6076, pp. 74-83, Feb. 2006.

[11] H. Liu, N. Klomp and I. Heynderickx, “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] H. R. Wu and M. Yuen, “A generalized block-edge impairment metric for video coding,” IEEE Signal Processing Letters, vol. 4, no. 11, pp. 317–320, 1997.

[13] H. Liu and I. Heynderickx, “A Perceptually Relevant No-Reference Blockiness Metric Based on Local Image Characteristics,” EURASIP Journal on Advances in Signal Processing, vol. 2009, 2009.

[14] K. Zon and W. Ali, “Automated video chain optimization,” IEEE Transactions on Consumer Electronics, vol. 47, pp. 593-603, Aug 2001.

[15] C. C. Koh, S. K. Mitra, J. M. Foley, and I. Heynderickx, “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, March 2005.

[16] I. O. Kirenko, R. Muijs, and L. Shao, “Coding artifact reduction using non-reference block grid visibility measure,” in Proc. IEEE International Conference on Multimedia and Expo, pp. 469–472, July 2006.

[17] Z. Wang, H. R. Sheikh and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images", in Proc. IEEE International Conference on Image Processing, vol. 1, pp. 477-480, September 2002.
[18] P. Marziliano, F. Dufax, S. Winkler and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to JPEG2000," Signal Processing: Image Communication, vol. 19, pp. 163-172, 2004.
[19] Z. Sazzad, Y. Kawayoke, and Y. Horita, "No reference image quality assessment for JPEG2000 based on spatial features," Signal Processing: Image Communication, vol. 23, pp. 257-268, 2008.
[20] H. R. Sheikh, A. C. Bovik, and L. K. Cormack, "No-Reference Quality Assessment Using Natural Scene Statistics: JPEG2000," IEEE Transactions on Image Processing, vol. 14, no. 11, pp. 1918-1927, December 2005.
[21] P. Gastaldo, S. Rovetta and R. Zunino, "Objective Quality Assessment of MPEG-2 Video Streams by Using CBP Neural Networks," IEEE Transactions on Neural Networks, vol. 13, no. 4, pp. 939-947, 2002.
[22] P. Gastaldo, R. Zunino, I. Heynderickx and E. Vicario, "Objective quality assessment of displayed images by using neural networks," Signal Processing: Image Communication, vol. 20, pp. 643-661, 2005.
[23] P. Gastaldo and R. Zunino, "Neural networks for the no-reference assessment of perceived quality," Journal of. Electronic Imaging, 14 (3), 033004, 2005.
[24] R. V. Babu, S. Suresh and A. Perkis, "No-reference JPEG-image quality assessment using GAP-RBF," Signal Processing, vol. 87, no.6, pp.1493-1503, 2007.
[25] M. Yuen and H. R. Wu, "A survey of hybrid MC/ DPCM/ DCT video coding distortions," Signal Processing, vol. 70, no. 3, pp. 247-278, November 1998.
[26] M. C. Q. Farias, M. S. Moore, J. M. Foley and S. K. Mitra, "Perceptual contributions of blocking, blurring, and fuzzy impairments to overall annoyance," in Proc. of SPIE, Human Vision and Electronic Imaging IX, vol. 5292, pp. 109-120, January 2004.
[27] J. Xia, Y. Shi, K. Teunissen and I. Heynderickx, "Perceivable artifacts in compressed video and their relation to video quality," Signal Processing: Image Communication, 2009.
[28] H. Liu, N. Klomp and I. Heynderickx, "A Perceptually Relevant Approach to Ringing Region Detection," IEEE Trans. Image Processing, June 2010.
[29] Redi, J., Gastaldo, P., Zunino, R. and Heynderickx, I., "Reduced reference assessment of perceived quality by exploiting color information", Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2009.
[30] D. E. Rumelhart and J. L. McClelland, Parallel distributed processing, MIT Press, Cambridge, MA, 1986.
[31] S. Ridella, S. Rovetta and R. Zunino, "Circular back-propagation networks for classification," IEEE Trans. on Neural Networks, vol. 8, pp. 84-97, 1997.
[32] E. B. Bau, and H. David, "What size net gives valid generalization?" Neural Comput., vol. 1, pp. 151-160, 1989.
[33] B. Widrow and M. A. Lehr, "30 Years of Adaptive Neural Networks: Perceptron, Madaline and Back Propagation," Proc. IEEE, 78(9), pp. 1415-42, 1990.

[34] T. P. Vogel, J. K. Mangis, A. K. Rigler, W. T. Zink and D. L. Alkon, “Accelerating the convergence of the back propagation method,” Biol. Cybern., vol. 59, pp. 257-263, 1988.
[35] 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
[36] Y. Liu, “Unbiased estimate of generalization error and model selection in neural network,” Neural Networks, vol. 8, pp215-219, 1995.
[37] L. Prechelt, “Automatic early stopping using cross validation: quantifying the criteria,” Neural Networks, vol. 4, pp761-767, 1998.
[38] R. Hecht-Nielsen, Neurocomputing, Addison-Wesley, Reading, MA, 1989.
[39] 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
[40] S. Winkler, “Vision Models and Quality Metrics for Image Processing Applications,” Ph.D. dissertation, Dept. Elect., EPFL, Lausanne, 2002.
[41] H. Liu, J. Redi, H. Alers, R. Zunino and I. Heynderickx, "No-reference image quality assessment based on localized gradient statistics: application to JPEG and JPEG2000", IS&T/ SPIE Electronic Imaging 2010, Human Vision and Electronic Imaging XV, January 2010.
[42] J. Redi, P. Gastaldo, I. Heynderickx, R. Zunino, “Color Distribution Information for the Reduced-Reference Assessment of Perceived Image Quality,” IEEE Trans. Circuits and Systems for Video Technology, 2010, in press.

Chapter 6

Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data

***Abstract*: Since the human visual system (HVS) is the ultimate assessor of image quality, current research on the design of objective image quality metrics tends to include an important feature of the HVS, namely visual attention. Different metrics for image quality prediction have been extended with a computational model of visual attention, but the resulting gain in reliability of the metrics so far was variable. To better understand the basic added value of including visual attention in the design of objective metrics, we used measured data of visual attention. To this end, we performed two eye-tracking experiments: one with a free-looking task and one with a quality assessment task. In the first experiment twenty observers looked freely to twenty-nine unimpaired original images, yielding us so-called natural scene saliency. In the second experiment twenty different observers assessed the quality of distorted versions of the original images. The resulting saliency maps showed some differences with the natural scene saliency, and therefore, we applied both types of saliency to four different objective metrics predicting the quality of JPEG compressed images. For both types of saliency the performance gain of the metrics improved, but to a larger extent when adding the natural scene saliency. As a consequence, we further integrated natural scene saliency in several state-of-the-art quality metrics, including three full-reference metrics and two no-reference metrics, and evaluated their prediction performance for a larger set of distortions. By doing so, we evaluated whether and to what extent the addition of natural scene saliency is beneficial to objective quality prediction in general terms. In addition, we address some practical issues in the design of an attention-based metric. The eye-tracking data are made available to the research community [1].**

This chapter is based on the research article accepted as "Visual Attention in Objective Image Quality Assessment: Based on Eye-Tracking Data" by H. Liu and I. Heynderickx for IEEE Transactions on Circuits and Systems for Video Technology.

6.1 Introduction

Image quality metrics are already integrated in a broad range of visual communication systems, for example for the optimization of digital imaging systems, the benchmarking of image and video coding algorithms, and the quality monitoring and control in displays [2]. These so-called *objective* metrics have the aim to automatically quantify the perceived image quality, and so, to serve eventually as an alternative for expensive quality evaluation by human observers. They range from dedicated metrics that measure a specific image distortion to general metrics that assess the overall perceived quality. Both the dedicated and general metrics can be classified into full-reference (FR), reduced-reference (RR), and no-reference (NR) metrics, depending on to what extent they use the original, non-degraded image or video as a reference. FR metrics are based on measuring the similarity between the distorted image and its original version. In real-world applications where the original is not available, RR and NR metrics are used. RR metrics make use of features extracted from the original, while NR metrics attempt to assess the overall quality or some aspect of it without the use of the original.

Since the human visual system (HVS) is the ultimate assessor of image quality, it is highly desirable to have objective metrics that predict image or video quality consistent with what humans perceive [2]. Traditional FR metrics, such as the mean squared error (MSE) or the peak signal-to-noise ratio (PSNR), are simple, since they are purely defined on a pixel-by-pixel difference between the distorted and the original image, but, they are also known for their poor correlation with perceived quality [3]. Therefore, a considerable amount of research is devoted to the development of more reliable objective metrics taking characteristics of the HVS into account.

Some meaningful progress in the design of HVS-based objective metrics is reported in the literature [4]-[18]. In these studies, lower level aspects of the HVS, such as contrast sensitivity, luminance masking and texture masking, are successfully modeled and integrated in various metrics. The basic idea behind the metrics in [4]-[7] is to decompose the image signal into channels of various frequencies and orientations in order to reflect human vision at the neural cell level. Classical HVS models, such as the contrast sensitivity function (CSF) per channel, and interactions between the channels to simulate masking, are then implemented. These metrics are claimed to be perceptually more meaningful than MSE or PSNR. In [8]-[13], metrics are designed to explicitly quantify the annoyance of various compression artifacts. In this research, properties of the HVS are combined with the specific physical characteristics of the artifacts to estimate their supra-threshold visibility to the human eye. The added value of including HVS aspects in these metrics is validated with psychovisual experiments. Instead of simulating the functional components of the HVS, the metrics in [14]-[18] are rather based on the overall functionality of the HVS, e.g. by assuming that the HVS separates structural information from nonstructural information in the scene [14]. These metrics are able to successfully predict image quality in close agreement with human judgments.

In recent years, researchers tend to include higher level aspects of the HVS, such as visual attention, in objective metrics. Limited progress has been made in this research area, mainly due to the fact that the mechanism of attention for image quality judgment is not fully understood yet, and also due to the difficulties of

precisely modeling visual attention. Current research mostly incorporates visual attention into the objective metrics in an ad-hoc way, based on optimizing the performance increase in predicting perceived quality. For example, studies in [19]-[23] are based on the assumption that a distortion occurring in an area that gets the viewer's attention is more annoying than in any other area, and they attempt to weight local distortions with local saliency, a process referred to as "visual importance pooling". The essential concept behind this approach is that the natural scene saliency (i.e. saliency driven by the original image content, and referred to as NSS) and the image distortions are taken into account separately, and they are combined to determine the overall quality score. In such a scenario, a variety of computational attention models are implemented in different metrics, resulting in a performance gain as reported in [19]-[23]. As such, this approach appears to be a viable way of including visual attention in objective metrics.

There are, however, several concerns related to the development of attention-based objective quality metrics. First of all, most research published so far in the literature employs an existing attention model to specifically optimize a targeted objective metric. Computational attention models are available, e.g. in [24] and [25], but they are either designed or chosen for a specific domain, and therefore, not necessarily generally applicable. Moreover, the accuracy of these models in predicting human visual attention is not always completely proved yet, especially not in the domain of image quality assessment. Therefore, the question arises whether an attention model successfully embedded in one particular metric is also able to enhance the performance of other metrics, and even if so, whether the gain by adding this attention model to a specific metric is comparable to the gain that can be obtained with alternative metrics. Secondly, it is well known that eye movements depend on the task assigned to the observer [26]. Hence, whether NSS or saliency during image quality assessment should be included in the design of objective quality metrics is still insufficiently studied. It is, e.g., not known yet whether the difference between both types of saliency is sufficiently large to actually affect the performance gain for the objective quality metrics. Thirdly, since computational efficiency becomes a significant issue when applying an objective metric in real-time processing, the measured gain in metric performance should be balanced against the additional costs needed for the rather complex attention modeling. This implies that before implementing an attention-based metric, it is worthwhile to know exactly whether and to what extent including visual attention can improve existing objective quality metrics. Finally, studies combining visual attention and image distortions in a perceptually meaningful way are still limited, and hardly discuss a generalized strategy for combining distortion visibility and saliency.

Obviously, investigating the aspects mentioned above heavily relies on the reliability of the visual attention data used. Since recording eye movements is so far the most reliable means for studying human visual attention [26], it is highly desirable to use these "ground truth" visual attention data for the evaluation of the added value of attention in objective quality metrics. This idea is recently exploited in [27], in which the data of an eye-tracking experiment are integrated in the PNSR and SSIM [14] metric. The results obtained in [27], however, are inconsistent with those found in [19]-[23], i.e. no clear improvement is found in the metric performance when weighting the local distortions with local saliency. It should,

however, be noted that the eye-tracking data of [27] were collected during image quality assessment with the DSIS (Double Stimulus Impairment Scale) protocol [28]. This implies that each observer saw an unimpaired reference and its impaired version several times during the experiment. As a consequence, the observer might have learnt where to look for the artifacts, and thus, the recorded eye-tracking data on the impaired images may have been more affected by the image distortions than by the natural scene content. Simply adding then these eye-tracking data to a quality metric may overweight the distraction power of the distortions compared to the NSS, and this may explain differences in the conclusions between [27] and [19]-[23]. To evaluate these assumptions, more data on whether to include NSS or saliency during scoring in the design of an attention-based metric is needed. This issue is addressed in [29] and [30], and the results show a trend of a larger improvement in predictability of the objective metrics when using eye-tracking data obtained during freely looking to unimpaired images. It should, however, be kept in mind that the study reported in [29] and [30] only made use of a limited number of human subjects (five participants looked freely to the images, while two scored the images). Nonetheless, the observed trend is in line with research recently published in [31], showing that adding "ground truth" NSS (in this case obtained by asking human observers to select the region-of-interest (ROI) in reference images) significantly improves the performance of metrics that predict the perceived quality of images that are wirelessly transmitted. Artifacts in these images are typically clustered in certain areas of the image. In such a specific scenario, using NSS is more practical since it can be transmitted as side information through the wireless communication channel. As such, the metric can make use of ROI versus background (BG) segmentation at the receiver end in real-time.

To better understand the added value of including visual attention in the design of objective metrics, we start from eye-tracking data obtained during free looking and during scoring image quality, as explained in Section II. Both types of saliency are then added to several objective quality metrics well-known in literature. The corresponding results are discussed in Section III, and reveal that although both types of saliency are beneficial for objective quality prediction, NSS tends to improve the metrics' performance more. As a consequence, we integrate, as discussed in Section IV, NSS in three full-reference metrics and two no-reference metrics with the aim to provide more accurate quantitative evidence on whether and to what extent visual attention can be beneficial for objective quality prediction. We also discuss some important issues of applying NSS in the design of an attention-based metric. Moreover, we have made the eye-tracking data publicly available [1] to facilitate future research in image quality assessment.

6.2 Eye-Tracking Experiments

It is generally agreed that under normal circumstances human eye movements are tightly coupled to visual attention [32]-[34]. Therefore, we performed eye-tracking experiments to obtain "ground truth" visual attention data. Actually, two eye-tracking experiments were conducted. In the first experiment, the NSS for the twenty-nine source images of the LIVE database [35] was collected by asking twenty observers to look freely to the images. In the second experiment, the

saliency was recorded for twenty different observers, who were requested to score the quality of distorted versions of the source images.

6.2.1 Test Environment

The eye-tracking experiment was carried out in the New Experience Lab of the Delft University of Technology [36]. Eye movements were recorded with an infrared video-based tracking system (iView X RED, SensoMotoric Instruments). It had a sampling rate of 50 Hz, a spatial resolution of 0.1°, and a gaze position accuracy of 0.5°-1.0°. Since the system could compensate for head movements within a certain range, a chin rest was sufficient to reduce head movements and ensure a constant viewing distance of 70cm. The stimuli were displayed on a 19-inch CRT monitor with a resolution of 1024x768 pixels and an active screen area of 365x275mm. Forty students, being twenty-four males and sixteen females, inexperienced with eye-tracking recordings, were recruited as participants. They were assigned to two groups of equal size (Group A and B), each with twelve males and eight females. Each session (per subject) was preceded by a 3x3 point grid calibration of the eye-tracking equipment.

6.2.2 Experiment I: NSS

Participants of Group A were requested to look freely to the twenty-nine source images of the LIVE database [35]. Each participant saw all stimuli in a random order. Each stimulus was shown for 10s followed by a mid-gray screen during 3s. The participants were requested to look at the images in a natural way ("view it as you normally would").

6.2.3 Experiment II: Saliency during Scoring

Participants of Group B were requested to score JPEG compressed versions of the source images (using MATLAB's *imwrite* function). To include a broad range of quality, while avoiding that the recorded saliency was biased by viewing a scene multiple times, the source images were divided into six groups (i.e. five groups of five scenes each, and one group of four scenes, indicated by "S1" to "S6"). Each group of scenes was compressed at a different level (i.e. S1 at Q=5, S2 at Q=10, S3 at Q=15, S4 at Q=20, S5 at Q=30, and S6 at Q=40). By doing so, each scene was viewed only once per subject, and for each subject in a different random order. The subject was requested to score the image quality for each stimulus with the single-stimulus (SS) method, i.e. in the absence of a reference [28]. A categorical scoring scale (recommended by ITU-R [28]) with the semantic terms "excellent", "good", "fair", "poor" and "bad" was used. Each stimulus was shown for 10s, followed by a scoring screen as illustrated in Figure 1. The actual experiment was preceded by a training, in which the participant was instructed on the task and could familiarize himself/ herself with how to use the scoring scale.

Fig. 1. Illustration of the scoring screen.

6.3 NSS versus Saliency during Scoring Applied in Objective Metrics

6.3.1 Saliency Map

A saliency map representative for visual attention is usually derived from the spatial pattern of fixations in the eye tracking data [32]-[34]. To construct this map, each fixation location gives rise to a gray-scale patch whose activity is Gaussian distributed. The width (σ) of the Gaussian patch approximates the size of the fovea (about 2° of visual angle). A mean saliency map (MSM) over all fixations of all subjects is then calculated as follows:

$$S_i(k,l) = \sum_{j=1}^{T} \exp[-\frac{(x_j - k)^2 + (y_j - l)^2}{\sigma^2}] \quad (1)$$

where $S_i(k, l)$ indicates the saliency map for stimulus I_i of size M×N pixels (i.e. $k\epsilon[1, M]$ and $l\epsilon[1, N]$), (x_j, y_j) are the spatial coordinates of the jth fixation (j=1…T), T is the total number of all fixations over all subjects, and σ indicates the standard deviation of the Gaussian (i.e. σ =45 pixels in our specific case). The intensity of the resulting saliency map is linearly normalized to the range [0, 1]. Figure 2 illustrates as an example a MSM derived from eye-tracking data obtained in experiment I for one of the original images, and the MSM obtained in experiment II for a JPEG compressed version of the same image (the saliency maps for the entire database can be accessed in [1]).

The example illustrates typical correspondences and differences between the NSS, derived from experiment I, and the saliency during scoring, derived from experiment II. In general, the most salient regions are comparable between the NSS and the saliency during scoring, but there are some deviations for which it is worthwhile to investigate their impact on the performance of an objective metric. An extensive discussion on the differences between NSS and saliency during scoring, including aspects of the appropriate comparison method, and the impact of

the experimental protocol, is outside the scope of this paper, and will be treated in a separate contribution [37].

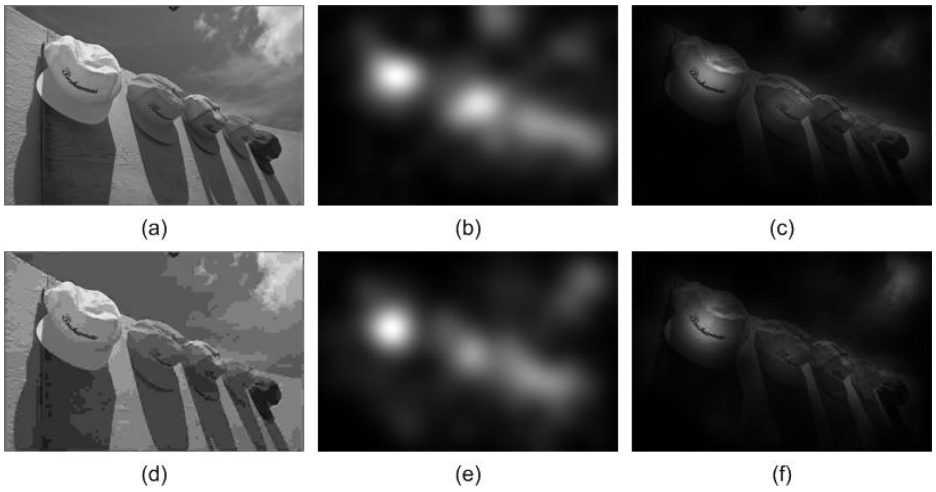


Fig. 2. Illustration of the saliency map: (a) original image, (b) mean saliency map (MSM) of (a) derived from the eye-tracking data of experiment I, (c) saliency map (b) superimposed on the image (a), (d) JPEG compressed image (Q=5), (e) mean saliency map (MSM) of (d) derived from the eye-tracking data of experiment II, (f) saliency map (e) superimposed on the image (d). Note that the darker the regions are, the lower the saliency is.

6.3.2 The Added Value of NSS and Saliency during Scoring in Objective Metrics

Based on the eye-tracking data, obtained from both our experiments, we evaluate whether and to what extent adding saliency is beneficial to the prediction performance of objective metrics. In this evaluation we compare the performance gain obtained when adding NSS versus saliency during scoring. To this end, we use the subjective scores we obtained in experiment II, and we try to predict these scores with several well-known objective metrics, all weighted with both types of saliency.

Subjective Scores

In experiment II, twenty human subjects scored the quality of twenty-nine JPEG distorted images. We transformed the raw quality ratings (i.e. “excellent”=5, “good”=4, “fair”=3, “poor”=2 and “bad”=1 as shown in Figure 1) into numbers, and calculated the Mean Opinion Score (MOS) as described in [13]. The resulting MOS are illustrated in Figure 3.

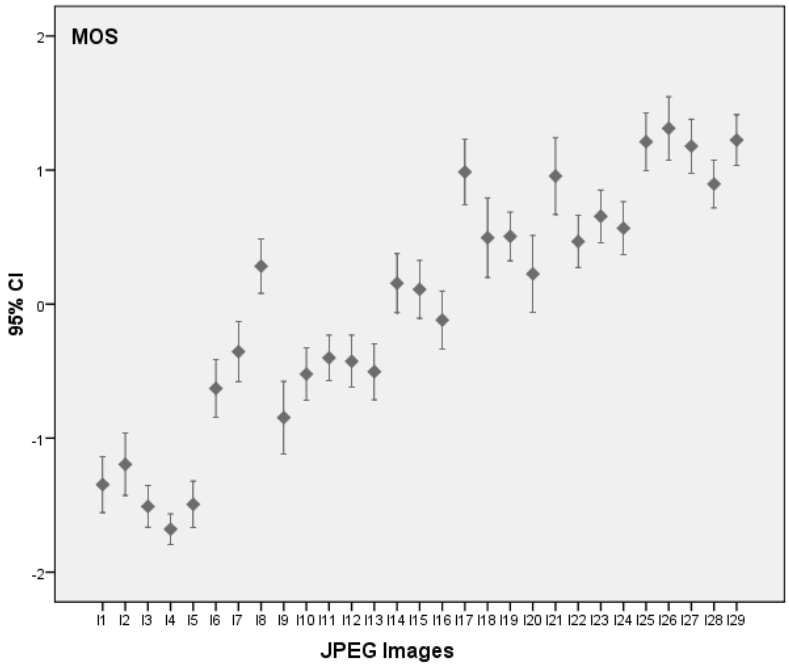


Fig. 3. The mean opinion scores (MOS) of the 29 JPEG images of experiment II. The error bars indicate the 95% confidence interval.

Objective Metrics

The evaluation of adding saliency was performed with four objective metrics (i.e. three FR metrics and one NR metric), which are so far widely accepted in the image quality community to assess the quality of JPEG compressed images. The FR metrics are:

- PSNR: The Peak Signal-to-Noise Ratio simply measures the difference (i.e. mean squared error) between the distorted image and its original version on a pixel-by-pixel base.
- SSIM: The Structural SIMilarity index [14] assumes that the HVS is highly adapted for extracting structural information from a scene, and it measures image quality based on the degradation in structural information.
- VIF: The Visual Information Fidelity [15] quantifies how much of the information present in the reference image can be extracted from the distorted image. Note that in this paper we use the implementation of the VIF in the spatial domain (as described in [35]).

The NR metric is:

- GBIM: The Generalized Block-edge Impairment Metric [8] is one of the most well-known metrics to quantify blocking artifacts in DCT coding. It measures blockiness as an inter-pixel difference across block boundaries (i.e. referred to as block-edges) scaled with a weighting function, which addresses luminance and texture masking of the HVS.

The objective metrics mentioned above are all formulated in the spatial domain. They estimate the image distortion locally, yielding a quantitative distortion map, which provides a spatially varying quality degradation profile. As an example, figure 4 (a) illustrates the distortion map calculated by SSIM for the JPEG compressed image of Figure 2 (d) (bit rate of 0.41 bbp). The intensity value of each pixel in the distortion map indicates the local degree of distortion, i.e. the lower the intensity, the larger the distortion is.

Including Saliency

Saliency (i.e. either NSS or saliency during scoring) is included in a metric by locally weighting the distortion map, as illustrated in Figure 4 (b) and (c) for the distortion map of SSIM weighted with NSS and saliency during scoring, respectively. Note that in the case of GBIM, the metric is calculated only around block-edges. As a result, weighting its distortion map with saliency actually gives more weight to the block-edges in the salient areas than in the non-salient areas.

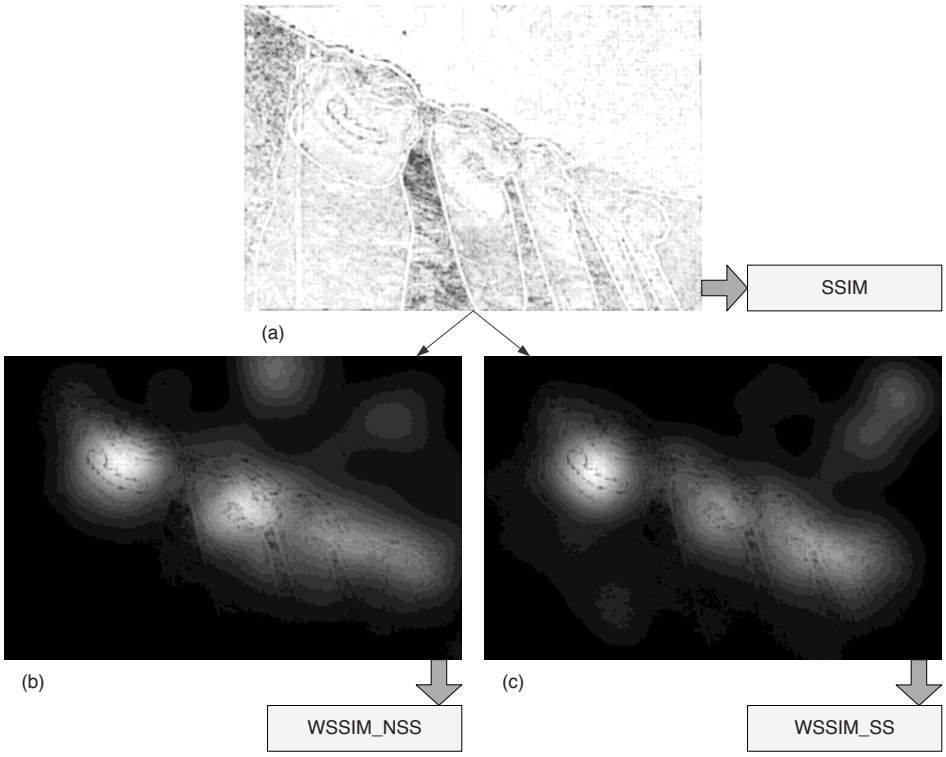


Fig. 4. Illustration of an objective metric based on saliency: (a) distortion map of SSIM calculated for the JPEG compressed image (bit rate 0.41bbp) of Fig 2 (d), (b) the corresponding NSS superimposed on (a), and (c) the corresponding saliency during scoring superimposed on (a). For the distortion map, the lower the intensity, the larger the distortion is.

Adding saliency to PSNR, SSIM, VIF and GBIM results in eight attention-based metrics, which are referred to as WPSNR_NSS, WPSNR_SS, WSSIM_NSS, WSSIM_SS, WVIF_NSS, WVIF_SS, WGBIM_NSS and WGBIM_SS, respectively. They can be defined as:

$$WMetric = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}[distortion\_map(x,y)\cdot S_i(x,y)]}{\sum_{x=1}^{M}\sum_{y=1}^{N}S_i(x,y)} \quad (2)$$

where *distortion_map* is calculated by the metric used, *S* indicates the corresponding saliency map derived from the eye-tracking experiment, and *WMetric* denotes the resulting attention based metric. It should be noted that the combination strategy used here is a simple weighting function similar to that in [19]-[23]. More complex combination strategies may further improve the metric's performance, as is discussed in Section IV.

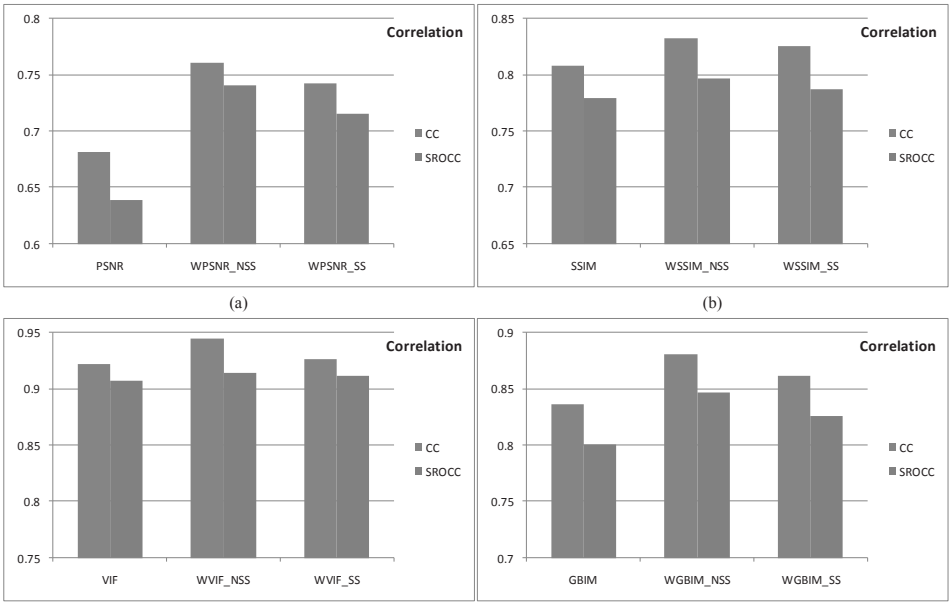
Experimental Results

As prescribed by the VQEG (Video Quality Experts Group) [38] the performance of an objective metric is determined by its ability to predict subjective quality ratings (the MOS). This ability can be quantified by the Pearson linear correlation coefficient (CC) indicating prediction accuracy, the Spearman rank order correlation coefficient (SROCC) indicating prediction monotonicity, and the root-mean-squared error (RMSE). With respect to the latter measure, we want to note that the scores are normalized to the scale [1, 10] before the calculation of the RMSE. As suggested in [38], the metric's performance can also be evaluated with non-linear correlations using a non-linear mapping of the objective predictions before computing the correlation. Indeed, the image quality community is more accustomed to e.g. a logistic function, to fit the predictions of an objective metric to the MOS. It may, for example, account for a possible saturation effect in the quality scores at high quality. A non-linear fitting usually yields higher correlation coefficients in absolute terms, while generally keeping the relative differences between the metrics [39]. On the other hand, without a sophisticated non-linear fitting (often including additional parameters) the correlation coefficients cannot mask a bad performance of the metric itself, as discussed in [23]. To better visualize differences in performance we avoid any non-linear fitting and directly use linear correlation and RMSE between the metrics' predictions and the MOS.

The twelve metrics (i.e. PSNR, WPSNR_NSS, WPSNR_SS, SSIM, WSSIM_NSS, WSSIM_SS, VIF, WVIF_NSS, WVIF_SS, GBIM, WGBIM_NSS and WGBIM_SS) are applied to the 29 JPEG compressed images, and the results are compared to the corresponding MOS of experiment II. Figure 5 shows the resulting CC, SROCC and RMSE-values, and demonstrates that the performance of all metrics enhances by including both NSS and saliency during scoring. The experimental results also tend to indicate that adding NSS to a metric yields a larger amount of performance gain than adding saliency during scoring. Adding NSS to PSNR corresponds to an increase of **8**% in CC and of **10**% in SROCC, and a decrease of 0.258 in the RMSE

value, but adding saliency during scoring to PSNR results only in an increase of **6**% in CC and of **8**% in SROCC, and a decrease of 0.225 in the RMSE value. The same trend of changes in performance is consistently found for the three other metrics.

Based on the above results, we can conclude that the small difference in saliency due to scoring with respect to the NSS is nonetheless sufficient to yield a consistent difference in performance gain when including visual attention to objective metrics. The relatively lower performance gain obtained with the saliency during scoring is possibly caused by the fact that this saliency is more spread towards background areas in the image due to the distraction power of annoying artifacts. As such, artifacts in background areas are weighted more (in relative terms) than artifacts in salient areas, and so, this might result in an overestimation of the annoyance of distortions in the background. Our results tend to support the assumption made in Section I for the difference in conclusion given in [27], on the one hand, and in [19]-[23], on the other hand. When adding saliency to objective metrics, it should be the NSS, obtained when people look at a distortion-free image for the first time. The saliency or distraction power of the image distortions themselves is kind of addressed by the metric (especially, when HVS aspects, such as contrast sensitivity and masking are already included in the distortion map).



	RMSE
PSNR	2.0269
WPSNR_NSS	**1.7688**
WPSNR_SS	1.8017

	RMSE
SSIM	2.01
WSSIM_NSS	**1.6923**
WSSIM_SS	1.8295

	RMSE
VIF	1.4697
WVIF_NSS	**0.9206**
WVIF_SS	1.0177

	RMSE
GBIM	1.465
WGBIM_NSS	**1.2687**
WGBIM_SS	1.3592

(e)

Fig. 5. Correlation coefficients and RMSE values (without nonlinear regression) of six metrics PSNR, WPSNR_NSS, WPSNR_SS, SSIM, WSSIM_NSS and WSSIM_SS for the 29 JPEG images of experiment II.

6.4 Adding NSS in Objective Metrics: Based on LIVE Database

To further evaluate the added value of visual attention in objective metrics, we include the NSS obtained from our eye-tracking data in experiment I into various objective metrics available in literature, and compare the performance of these attention-based metrics to the performance of the same metrics without visual attention. To also evaluate a variety of distortion types, this validation is done for the entire LIVE database [35], which consists of 779 images distorted with JPEG compression (i.e. JPEG), JPEG2000 compression (i.e. JP2K), white noise (i.e. WN), Gaussian blur (i.e. GBLUR), and simulated fast fading Rayleigh occurring in (wireless) channels (i.e. FF). Per image the database also gives a difference in mean opinion score (DMOS) derived from an extensive subjective quality assessment study [40]. Based on the evaluation, we address some technical issues relevant to the application of visual attention in objective metrics. More specifically, we discuss the effect of image content and of the combination strategy.

6.4.1 Objective Metrics

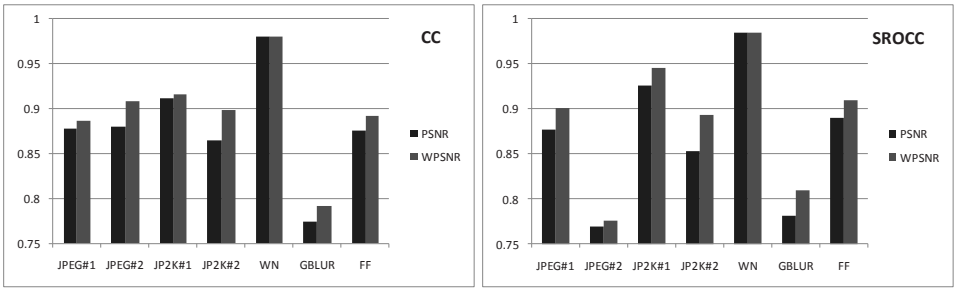
For practical reasons the objective metrics used in our validation are limited to three well-known FR metrics and two NR metrics. The FR metrics are PSNR, SSIM and VIF, as explained in Section III. The NR metrics are GBIM (also explained in Section III) and NRPB. The latter refers to the No-Reference Perceptual Blur metric [11] based on extracting sharp edges in an image, and measuring the width of these edges.

6.4.2 Evaluation of the Overall Performance Gain

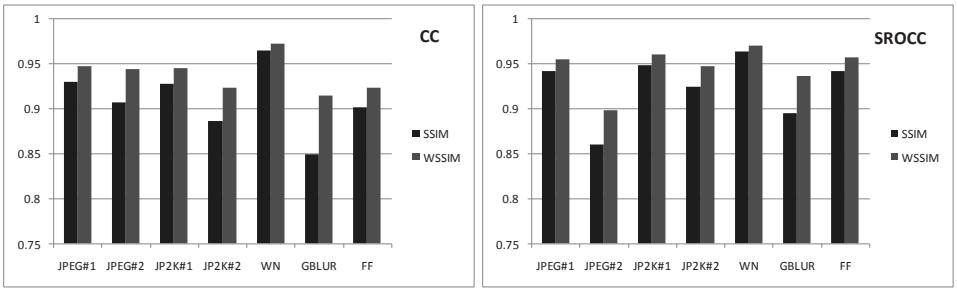
Adding NSS to the metrics mentioned above results in five attention-based metrics, which are referred to as WPSNR, WSSIM, WVIF, WGBIM and WNRPB, respectively. The six FR metrics, i.e. PSNR, SSIM, VIF, WPSNR, WSSIM and WVIF, are intended to assess image quality independent of distortion type, and therefore, are applied to the entire LIVE database [35]. The metrics GBIM and WGBIM are designed specifically for block-based DCT compression, and are applied to the JPEG#1 and JPEG#2 sub-sets of the LIVE database. The metrics NRPB and WNRPB are designed to quantify blur in images, and they are applied to the Gaussian blur sub-set of the LIVE database.

Figures 6 and 7 give the corresponding correlation coefficients and RMSE values. The overall gain (averaged over artifacts where appropriate) of an attention-based metric over its corresponding metric without NSS is summarized in Tables I and II. Both figures and tables demonstrate that there is indeed a gain in performance when including visual attention in the objective metrics PSNR, SSIM, VIF, GBIM and NRPB, independent of the metric used and of the image distortion type tested. The actual amount of performance gain, however, depends on the metric and on the distortion type. A promising performance gain (expressed in terms of CC) is found for the subset of the LIVE database distorted by Gaussian blur: the gain of WPSNR over PSNR is **2**%, of WSSIM over SSIM is **7**%, of WVIF over VIF is **2**%, and of WNRPB over NRPB is **5**%. The amount of performance gain, however, is relatively

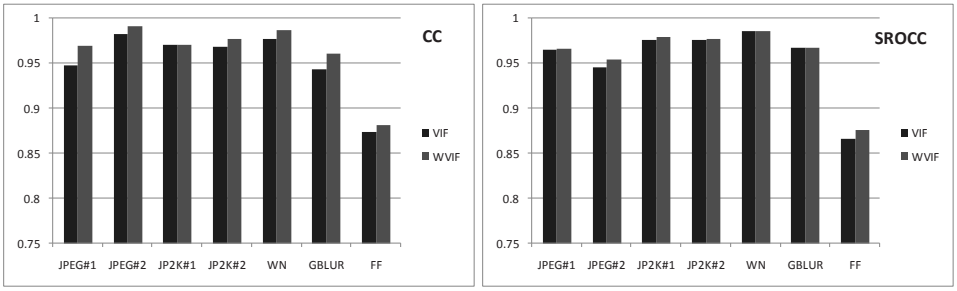
small for the subset of the LIVE database distorted by white noise: the gain (again in terms of CC) of WPSNR over PSNR is **0.01**%, of WSSIM over SSIM is **1**%, and of WVIF over VIF is **1**%. Differences in performance may be attributed to two possible causes: (1) the performance of a metric (i.e. without NSS) varies with the distortion type, and as such it is more difficult to obtain a significant increase in performance by adding NSS when a metric already has a high prediction performance for a given type of distortion, and (2) in the specific case of images distorted by Gaussian blur, some metrics might confuse unintended (Gaussian) blur with intended blur in the background to increase the field of depth (i.e. a high-quality foreground object with an intentionally blurred background). Adding NSS reduces the importance of blur in the background, and as such might improve the overall prediction performance of a metric.



(a) PSNR vs. WPSNR



(b) SSIM vs. WSSIM

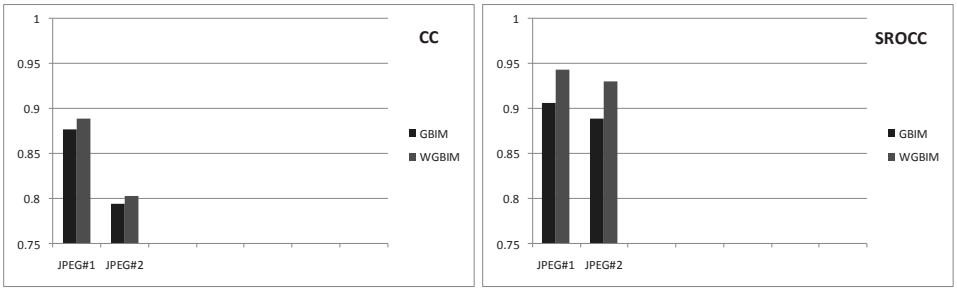


(c) VIF vs. WVIF

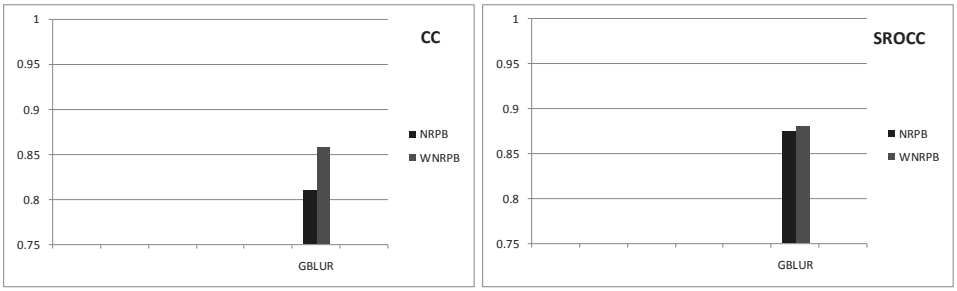
	JPEG#1	JPEG#2	JP2K#1	JP2K#2	WN	GBLUR	FF
PSNR	1.0586	1.2908	0.9926	1.3382	0.4422	1.3580	1.1585
WPSNR	**0.9853**	**1.1168**	**0.9050**	**1.1341**	**0.4383**	**1.2881**	**1.0578**
SSIM	0.7820	0.7540	0.8383	1.0441	0.6290	1.0666	0.9046
WSSIM	**0.6966**	**0.6317**	**0.7307**	**0.8479**	**0.6000**	**0.8282**	**0.8406**
VIF	0.6881	0.4881	0.5510	0.6563	0.5693	0.6762	1.2371
WVIF	**0.5882**	**0.4069**	**0.4964**	**0.5853**	**0.4364**	**0.6350**	**1.1933**

(d)

Fig. 6. Correlation coefficients and RMSE values (without nonlinear regression) of six full-reference (FR) metrics PSNR, WPSNR, SSIM, WSSIM, VIF and WVIF for images distorted by JPEG#1, JPEG#2, JPEG2000#1, JPEG2000#2, white noise (i.e. WN), Gaussian blur (i.e. GBLUR), and fast-fading (i.e. FF), respectively. Note that the data used are taken from the LIVE database [35].



(a) GBIM vs. WGBIM



(b) NRPB vs. WNRPB

	JPEG#1	JPEG#2
GBIM	0.9961	0.8144
WGBIM	**0.8108**	**0.6749**

	GBLUR
NRPB	1.0426
WNRPB	**0.9882**

(c)

Fig. 7. Correlation coefficients (without nonlinear regression) of four no-reference (NR) metrics: (a) GBIM and WGBIM for JPEG#1, JPEG#2, and (b) NRPB and WNRPB for Gaussian blur (i.e. GBLUR). The corresponding RMSE-values are given in (c).

TABLE I
Performance of PSNR, WPSNR, SSIM, WSSIM, VIF and WVIF averaged over all distortion types for the images of the LIVE database [35]

	CC	SROCC	RMSE
PSNR	0.88	0.87	1.09
WPSNR	0.90	0.90	0.99
Δ	ΔP=2%	ΔS=3%	ΔR=0.1

	CC	SROCC	RMSE
SSIM	0.91	0.92	0.86
WSSIM	0.94	0.95	0.74
Δ	ΔP=3%	ΔS=3%	ΔR=0.12

	CC	SROCC	RMSE
VIF	0.95	0.955	0.70
WVIF	0.96	0.958	0.62
Δ	ΔP=1%	ΔS=0.3%	ΔR=0.08

TABLE II
Performance of GBIM and WGBIM for the sub-sets JPEG#1 and JPEG#2, and performance of NRPB and WNRPB for the sub-set GBLUR of the LIVE database [35]

	CC	SROCC	RMSE
GBIM	0.83	0.90	0.91
WGBIM	0.84	0.94	0.74
Δ	ΔP=1%	ΔS=4%	ΔR=0.17

	CC	SROCC	RMSE
NRPB	0.81	0.87	1.04
WNRPB	0.86	0.88	0.99
Δ	ΔP=5%	ΔS=1%	ΔR=0.05

6.4.3 Statistical Significance

In order to check whether the numerical difference in performance between a metric with NSS and the same metric without NSS is statistically significant, we performed some hypothesis testing to provide statistical soundness on the conclusion of superiority of the attention-based metrics. As suggested in [38], the test is based on the residuals between the DMOS and the quality predicted by the metric (hereafter referred to as M-DMOS residuals). Before being able to do a parametric test, we evaluated the assumption of normality of the M-DMOS residuals. A simple Kurtosis-based criterion (as used in [40]) was used for normality: if the residuals had a kurtosis between 2 and 4, they were assumed to be normally distributed, and the difference between the two sets of M-DMOS residuals could be tested with a parametric test. The results of the test for normality are summarized in Table III, and indicate that in most cases the residuals are normally distributed. Considering that most parametric tests are not too sensitive to deviations from normality, we decided to test statistical significance for the performance improvement of NSS based metrics with a parametric test for all combinations of objective metrics with distortion types. In our particular case, the two sets of residuals being compared are dependent samples: one is from the metric itself, and one is from the same metric after adding the NSS. Therefore, a paired-sample *t*-test [41] is used instead of the *F*-test, as suggested in [38], since the latter one assumes that the two samples being compared are independent. The paired-sample *t*-test starts from the null hypothesis stating that the residuals of one metric are statistically indistinguishable (with 95% confidence) from the residuals of that same metric with NSS. The results of this *t*-test are given in Table IV for all metrics and

distortion types separately. This table illustrates that in most cases the improvement in prediction performance by adding NSS to an objective metric is statistically significant. The improvement reported in Section IV.B is not statistically significant only in three combinations of metrics applied to a given distortion type (with only 29 stimuli).

TABLE III
Normality of the M-DMOS residuals: "1" means that the residuals can be assumed to have a normal distribution since the Kurtosis lies between 2 and 4.

	JPEG#1	JPEG#2	JP2K#1	JP2K#2	WN	GBLUR	FF
PSNR	1	1	1	1	1	1	1
WPSNR	1	1	1	1	1	1	1
SSIM	1	1	1	1	1	0	1
WSSIM	1	1	1	1	1	0	1
VIF	1	1	1	1	1	1	1
WVIF	1	1	1	1	1	1	1
GBIM	1	0					
WGBIM	1	0					
NRPB						1	
WNRPB						1	

TABLE IV
Results of *t*-test based on M-DMOS residuals: "1" means that the attention-based metric is statistically significantly better that the metric without NSS, and "-" means that the difference is not statistically significant.

	JPEG#1	JPEG#2	JP2K#1	JP2K#2	WN	GBLUR	FF
PSNR&WPSNR	1	1	1	-	1	1	-
SSIM&WSSIM	1	1	1	1	1	1	1
VIF&WVIF	1	1	1	1	1	1	1
GBIM&WGBIM	1	-					
NRPB&WNRPB						1	

It should, however, be noted that statistical significance testing is not straightforward, and the conclusions drawn from it largely depend e.g. on the number of sample points, on the selection of the confidence criterion, and on the assumption of normality of the residuals. These issues are extensively discussed in [40].

6.4.4 Evaluation of the Influence of Image Content

The distribution of saliency over an image largely depends on its content, and therefore it makes sense to also study whether the added value of including visual attention to objective metrics is content dependent. The effect of content on NSS is quantified by calculating per image the correlation between the MSM obtained from experiment I and each individual saliency map (ISM) (derived from the fixations of an individual subject). The correlation between two saliency maps (i.e. SM_A and SM_B) is often measured by the coefficient (ρ), as employed in [32]. It is defined as follows, with its value ranging between [-1, 1]:

$$\rho = \frac{\sum_{n=1}^{M} (SM_A(n) - \mu_A)(SM_B(n) - \mu_B)}{\sqrt{\sum_{n=1}^{M} (SM_A(n) - \mu_A)^2 \sum_{n=1}^{M} (SM_B(n) - \mu_B)^2}} \quad (3)$$

where μ_A and μ_B are the mean values of the SM_A and SM_B, respectively. M is the total number of pixels in both maps. A higher value of ρ indicates a larger similarity between the two saliency maps. Figure 8 gives the ρ -values between the MSM and the ISM averaged over all subjects. This averaged ρ -value strongly varies over the different natural scenes, with the highest value of ρ for "scene25" (ρ =0.7549) and the lowest value of ρ for "scene3" (ρ =0.4521). This averaged ρ -value quantifies the variation in eye-tracking behavior among human subjects when viewing a single stimulus. A large value of the ρ averaged over all subjects indicates a small variation in saliency among subjects, while a small value of ρ indicates that the saliency is widely spread among subjects. Figure 9 presents the images with the three smallest values of the averaged ρ (i.e. "set_low") in Figure 8. These images clearly lack highly salient features, and their corresponding MSM includes fixations distributed all over the image. Figure 10 shows the three images, with the largest value of the averaged ρ (i.e. "set_high") in Figure 8. These images generally contain a few salient features, such as the human face in the images "statue" and "studentsculpture" and the billboard in the image "cemetry". For these images the saliency converges around these features in the MSM. The difference in saliency between both sets of images is apparently driven by image content.

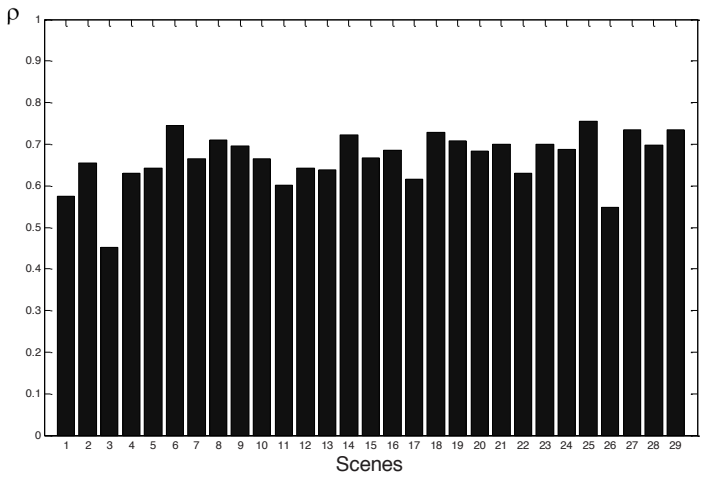


Fig. 8. The correlation coefficient (ρ) between the mean saliency map (MSM) and the individual saliency map (ISM) averaged over all subjects per scene.

To evaluate the content dependency in the performance gain when adding saliency to objective metrics, we repeated the experiment in Section IV.B once for the source images of "set_low", and once for the source images of "set_high". The former set contained 20 stimuli with JPEG compression, 17 stimuli with JPEG2000

compression, 15 stimuli with white noise, 15 Gaussian blurred stimuli, and 15 stimuli with fast fading artifacts, while the latter set consisted of 18 stimuli with JPEG compression, 17 stimuli with JPEG2000 compression, 15 stimuli with white noise, 15 Gaussian blurred stimuli, and 15 stimuli with fast fading artifacts. Figure 11 illustrates the comparison in performance gain (i.e. quantified by the Pearson correlation coefficient) between a metric and its NSS weighted version for the "set_low" and "set_high" images separately. In general, it shows the consistent trend that including saliency results in a larger performance gain in the objective metrics for the images of "set_high" than for the images of "set_low"; more particularly, for the images of "set_low", the performance gain when adding saliency is actually non-existing. The gain of WPSNR over PSNR corresponds to an average increase in the Pearson correlation coefficient (over all distortion types of the LIVE database) from 0.942 to 0.943 for the "set_low" images (i.e. **0.1**%), and from 0.882 to 0.910 for the "set_high" images (i.e. **2.8**%). The gain of WSSIM over SSIM is **0** (from 0.976 to 0.976) for the "set_low" images and **3.1**% (from 0.934 to 0.965) for the "set_high" images. The gain of WVIF over VIF is **0** (from 0.958 to 0.958) for the "set_low" images and **1.6**% (from 0.966 to 0.982) for the "set_high" images. The gain of WGBIM over GBIM is **1.6**% (from 0.929 to 0.945) for the "set_low" images and **7.7**% (from 0.789 to 0.866) for the "set_high" images. There is, however, one exception to this trend, namely for the metrics WNRPB and NRPB. As shown in Figure 11 (e), adding saliency degrades the performance of NRPB for the images of "set_high". This may be due to the specific design of the blur metric, which is based on measuring the width of extracted strong edges. Including the saliency of Figure 10 to the NRPB metric with a linear weighting combination strategy runs the risk of eliminating some very obvious edges in the calculation of blur, and may consequently affect the accuracy of the metric.

In summary, our findings suggest that the performance gain in an objective metric when applying saliency depends on the image content as well as on the specific metric design.

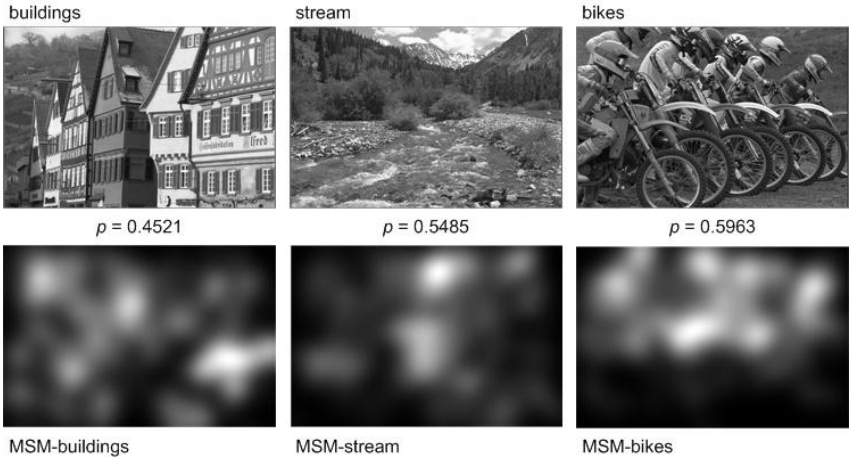


Fig. 9. Illustration of the three images with the smallest correspondence in saliency between subjects (i.e. smallest value of averaged ρ in Figure 8).

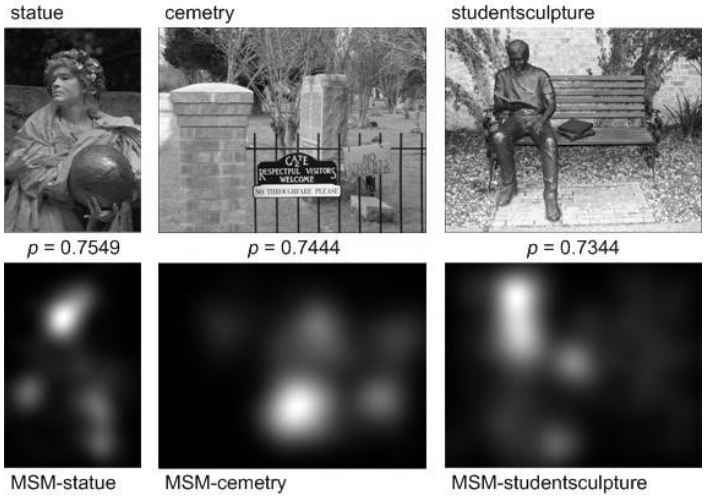
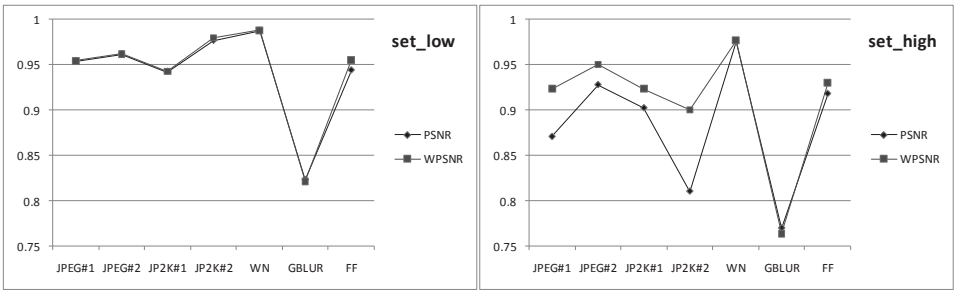
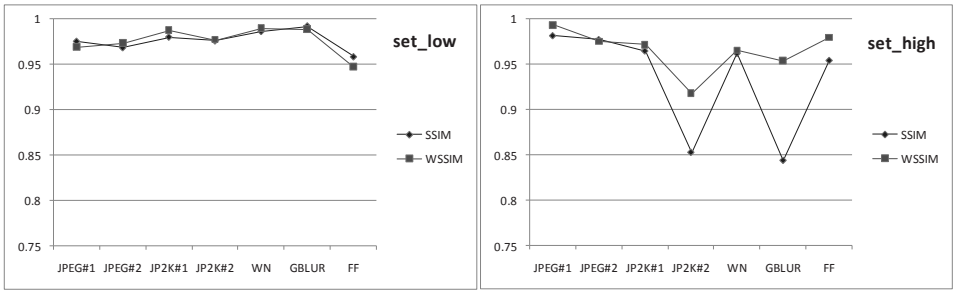


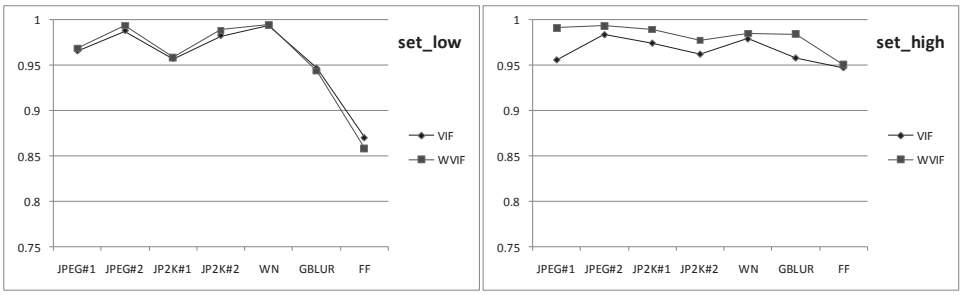
Fig. 10. Illustration of the three images with the largest correspondence in saliency between subjects (i.e. largest values of the averaged ρ in Figure 8).



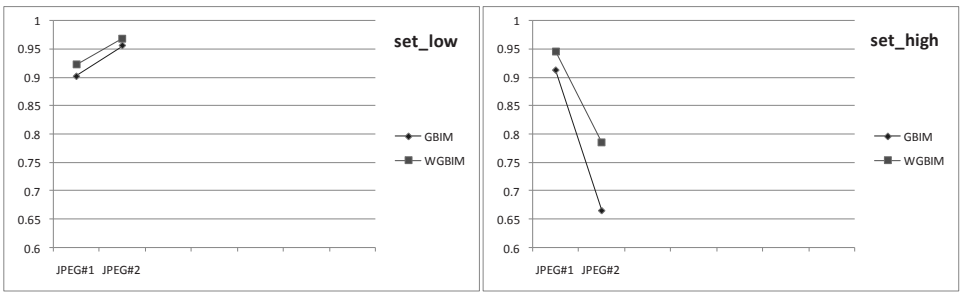
(a) PSNR vs. WPSNR



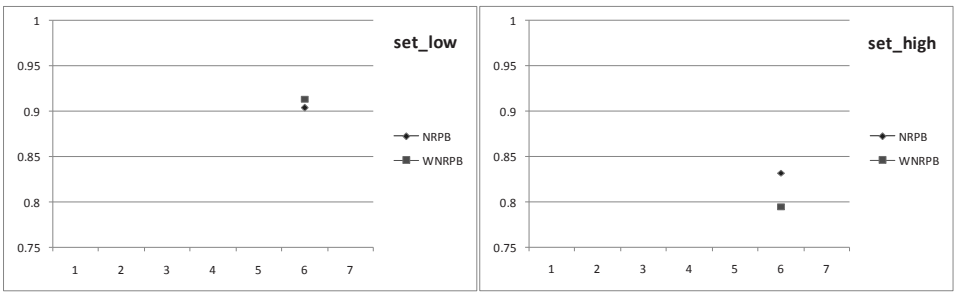
(b) SSIM vs. WSSIM



(c) VIF vs. WVIF



(d) GBIM vs. WGBIM



(e) NRPB vs. WNRPB

Fig. 11. Comparison in performance gain when adding saliency (quantified by the Pearson correlation coefficient) between images of "set_low" (distorted images extracted from the LIVE database [35] based on the source images of Figure 9) and images of "set_high" (distorted images extracted from the LIVE database [35] based on the source images of Figure 10).

6.4.5 Evaluation of the Influence of Combination Strategy

Fig. 12. An image JPEG compressed at a bit rate of 0.43bpp, and its corresponding NSS obtained from our eye-tracking data.

So far, saliency was added to the objective metrics based on a linear weighting combination strategy. This method is simple and intuitive, and has been widely adopted to pool local distortions of an image with saliency [19]-[23]. Our results of Section III and IV demonstrate the general effectiveness of using the linear combination strategy. This strategy, however, has limitations in dealing with certain distortions in more demanding conditions [42]. Figure 12 illustrates an image JPEG compressed at a bit rate of 0.43 bpp, and its corresponding NSS obtained from our eye-tracking data. Due to texture and luminance masking in the HVS [10], this image exhibits imperceptible blocking artifacts in the more salient areas (e.g. the foreground of the white tower), and relatively annoying blocking artifacts in the less salient areas (e.g. the background of the sky). In such a case, combining the distortion and saliency map with a linear combination strategy intrinsically underestimates the annoyance of the artifacts in the background, and their impact on the quality judgment.

To quantify the effect of linearly adding saliency in an objective metric for the quality prediction of demanding images, a subset of nine images was selected from the LIVE database. The images "img{9, 37, 44, 47, 63, 69, 89, 92, 105}" of the subset JPEG#1 typically represent the type of JPEG compressed images with the artifacts in the more salient areas locally masked by the content, and with clearly visible artifacts in the less salient areas. The blockiness metrics, GBIM and WGBIM are

applied to this sub-selection of the database. As illustrated in Figure 13, WGBIM fails in accurately predicting the subjective quality ratings for this subset of demanding images, mainly due to the inappropriate integration of saliency in the blockiness metric (i.e. the gain of WGBIM over GBIM in CC is **-59**%). Hence, the overall gain in CC of WGBIM over GBIM (i.e. **1**%) for the entire LIVE database of JPEG compressed images is explained by the fact that most of the images in this database exist of one of the following types: (1) images having visible artifacts uniformly distributed over the entire image, and (2) images having the artifacts masked by the content in the less salient areas, but showing visible artifacts in the more salient areas. Obviously, for these two types of images, adding saliency with a linear combination strategy is reasonable.

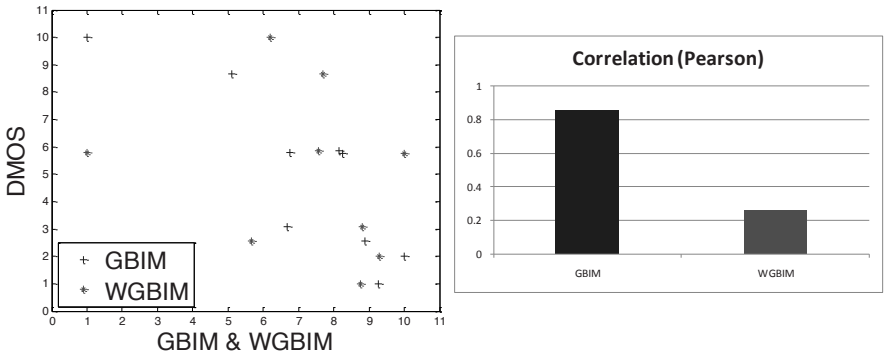


Fig. 13. Performance of the blockiness metrics GBIM and WGBIM in predicting the subjective quality rating of a subset of demanding images (i.e. img{9, 37, 44, 47, 63, 69, 89, 92, 105}) selected from the LIVE database JPEG#1 [35].

So, these findings indicate that a linear combination strategy is not necessarily appropriate for adding saliency in objective metrics. Hence, from a point of view of metric optimization, it is worthwhile to investigate adaptive combination strategies as, e.g., discussed in [23] and [42].

6.5 Discussion

In this paper we evaluate the intrinsic gain in prediction accuracy that can be obtained by introducing visual attention in objective quality metrics. This evaluation is performed for a diverse, though limited set of images, and mainly for distortions that affect the images globally. The results we obtained show that there is added value in weighting pixel-based distortion maps with local saliency. The amount of added value is bigger when extending the objective metrics with natural scene saliency than with saliency recorded while the viewers assess the quality of the images. The actual gain in performance accuracy is highly dependent on the image content, on the distortion type and on the objective metric itself. Images with a clear ROI demonstrate a bigger gain as compared to images in which the NSS is spread over the whole image. In addition, the gain is small for objective metrics that already show a high correlation with perceived quality for a given distortion type.

Although showing clear results, the study reported here has some limitations. First, as mentioned above, the set of images used has a fair size, but could be extended in order to investigate the effect of image content on the gain in prediction accuracy in a more systematic way. Second, most images are degraded with distortions that affect the image quality globally, i.e. the artifacts are uniformly distributed over the entire image. In specific applications, such as in wireless imaging, artifacts may occur localized, i.e. only at some random, but limited location in the image. Although we did not investigate this type of distortions specifically, we expect that introducing visual saliency in quality prediction metrics for this type of distortions is still beneficial. At least, results reported in [31] support this hypothesis. Finally, the gain in prediction accuracy claimed in this paper is based on eye-tracking recordings. These recordings intrinsically have some inaccuracy, which may limit the overall reliability of our conclusions. We have shown, however, that recorded saliency data are highly consistent when using well-calibrated equipment and a well-defined protocol; the consistency is even shown for data collected in various laboratories [43]. Using eye-tracking data, of course, is unrealistic for real-time applications. Hence, a visual attention model will be needed in the actual implementation of an objective metric. Since the reliability of most visual attention models is still limited, we expect that the actual gain in prediction accuracy that can be obtained in a real-time application is lower than what we showed here, at least with the current soundness of visual attention models. In the coming years the soundness of visual attention models may improve, but most probably at the expense of their computational cost.

Given the fact that the added value of having NSS weighted objective quality metrics depends on the image content, distortion type and objective metric, an adaptive approach might be desirable in real-time applications to limit the overall computational cost. In such an approach, the performance of an objective metric needed in the video chain can be optimized off-line; i.e. for each metric the added value of incorporating saliency can be estimated from its general prediction accuracy. For those metrics that contain saliency in their extended version a simple visual attention model can be used to determine the size of the ROI in the image. Only when the ROI is limited in size, the extended version of the metric is needed. Otherwise, the metric without saliency model can be applied at sufficient accuracy.

6.6 Conclusions

In this paper, we investigate the added value of visual attention in the design of objective metrics. Instead of using a computational model for visual attention, we conducted eye-tracking experiments to obtain “ground truth” visual attention data, thus making the results independent of the reliability of an attention model. Actually, two eye-tracking experiments were performed: one in which the participants looked freely to undistorted images, and a second one, in which different participants were asked to score the quality of a JPEG compressed version of the images. The resulting eye-tracking data indicate that there is some deviation between the natural scene saliency (NSS) and saliency during scoring.

Adding either type of saliency to an objective metric improves its performance in predicting perceived image quality. However, we also found a tendency that adding NSS to a metric yields a larger amount of gain in the performance. Based on

this evidence, the data of NSS were further integrated in several objective metrics available in literature, including three FR metrics and two NR metrics. This evaluation shows that there is indeed a gain in the performance for all these metrics when linearly weighting the local distortion map of the metrics with the NSS. The extent of the performance gain tends to depend on the specific objective metric and the image content. But our findings also illustrate that for some image content and for some distortion types, the linear combination strategy is insufficient and adaptive strategies are needed. Current and future research includes modeling saliency for real-time quality assessment, and integrating this saliency in objective metrics in a perceptually even more meaningful way.

6.7 References

[1] (2010) H. Liu and I. Heynderickx, "TUD Image Quality Database: Eye-Tracking Release 1", http:/ / mmi.tudelft.nl/ iqlab/ eye_tracking_1.html.

[2] Z. Wang and A. C. Bovik, Modern Image Quality Assessment. Synthesis Lectures on Image, Video, & Multimedia Processing, Morgan & Claypool, San Refael, Calif, USA, 2006.

[3] Z. Wang and A. C. Bovik, "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, Jan. 2009.

[4] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in A. B. Watson (Ed.), Digital images and human vision, pp. 179-206, The MIT Press, Cambridge, MA, 1993.

[5] J. Lubin, The use of psychophysical data and models in the analysis of display system performance. In A. B. Watson (Ed.), Digital Images and Human Vision, pp. 163–178. The MIT Press, Cambridge, MA, 1993.

[6] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," In Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, pp. 1945–1948, May 1989.

[7] A. B. Watson, J. Hu and J. F. Mcgowan, "DVQ: A digital video quality metric based on human vision," Journal of Electronic Imaging, 10(1), 20-29, 2001.

[8] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," IEEE Signal Processing Letters, vol. 4, no. 11, pp. 317–320, 1997.

[9] Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in Proc. IEEE International Conference on Image Processing, vol. 3, pp. 981-984, Sept. 2000.

[10] H. Liu and I. Heynderickx, "A Perceptually Relevant No-Reference Blockiness Metric Based on Local Image Characteristics," EURASIP Journal on Advances in Signal Processing, vol. 2009, 2009.

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

[12] R. Ferzli and L. J. Karam, "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.

[13] H. Liu, N. Klomp and I. Heynderickx, “A No-Reference Metric for Perceived Ringing Artifacts in Images,” IEEE Trans. on Circuits and Systems for Video Technology, 2010.
[14] 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, April 2004.
[15] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," IEEE Transactions on Image Processing, vol.15, no.2, pp. 430- 444, Feb. 2006.
[16] H. R. Sheikh, A. C. Bovik, and L. K. Cormack, "No-Reference Quality Assessment Using Natural Scene Statistics: JPEG2000," IEEE Transactions on Image Processing, vol. 14, no. 11, pp. 1918-1927, December 2005.
[17] R. V. Babu, S. Suresh and A. Perkis, “No-reference JPEG-image quality assessment using GAP-RBF,” Signal Processing, vol. 87, no.6, pp.1493-1503, 2007.
[18] P. Gastaldo and R. Zunino, “Neural networks for the no-reference assessment of perceived quality,” Journal of. Electronic Imaging, 14 (3), 033004, 2005.
[19] R. Barland and A. Saadane, “Blind Quality Metric using a Perceptual Importance Map for JPEG-2000 Compressed Images,” in Proc. IEEE Int. Conf. ICIP, pp. 2941-2944, Oct. 2006.
[20] D. V. Rao, N. Sudhakar, I. R. Babu and L. P. Reddy, “Image Quality Assessment Complemented with Visual Region of Interest,” in Proc. International conference on Computing: Theory and Applications, pp. 681-687, 2007.
[21] Q. Ma and L. Zhang, “Image quality assessment with visual attention,” in Proc. ICPR, pp. 1-4, Dec. 2008.
[22] N. G. Sadaka, L. J. Karam, R. Ferzli, and G. P. Abousleman, “A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling,” in Proc. IEEE Int. Conf. ICIP, pp. 369-372, Oct. 2008.
[23] A. K. Moorthy and A. C. Bovik, “Visual Importance Pooling for Image Quality Assessment,” IEEE Journal of Selected Topics in Signal Processing, Special Issue on Visual Media Quality Assessment, vol. 3, no.2, April 2009.
[24] L. Itti, C. Koch and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254-1259, Nov. 1998.
[25] U. Rajashekar, A. C. Bovik and L. K. Cormack, “Gaffe: A gaze-attentive fixation finding engine,” IEEE Trans. Image Processing, vol. 17, no. 4, pp 564-573, April 2008.
[26] A. L. Yarbus. Eye movements and vision. New York: Plenum Press, 1967.
[27] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, “Does where you Gaze on an Image Affect your Perception of Quality? Applying Visual Attention to Image Quality Metric,” in Proc. IEEE Int. Conf. ICIP, pp. 169-172, Oct. 2007.
[28] ITU-R Recommendation BT.500-11, Methodology for the subjective assessment of the quality of television pictures, International Telecommunication Union, Geneva, Switzerland, 2002.
[29] C. T. Vu, E. C. Larson, and D. M. Chandler, “Visual Fixation Patterns when Judging Image Quality: Effects of Distortion Type, Amount, and Subject Experience,” in Proc. IEEE SSIAI, pp. 73-76, March. 2008.

[30] E. C. Larson, C. T. Vu, and D. M. Chandler, "Can Visual Fixation Patterns Improve Image Fidelity Assessment?" in Proc. Intl. Conf. on Image Processing, pp. 2572-2575, October 2008.
[31] U. Engelke and H.-J. Zepernick, “Framework for optimal region of interest-based quality assessment in wireless imaging,” Journal of Electronic Imaging, vol. 19, no.1, ID 011005, 2010.
[32] N. Ouerhani, R. V. Wartburg, H. Hugli, and R. Muri, “Empirical Validation of the Saliency-based Model of Visual Attention,” Electronic Letters on Computer Vision and Image Analysis, 3(1): 13-24, 2004.
[33] D.D. Salvucci, A model of eye movements and visual attention, in: Third Internat. Conf. on Cognitive Modeling, pp. 252–259, 2000.
[34] C. Privitera, L. Stark, Algorithms for defining visual regions-of-interest: comparison with eye fixations, Pattern Anal. Mach. Intell. 22 (9), 970–981, 2000.
[35] 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
[36] New Delft Experience Lab. [Online]. Available: http:/ / mmi.tudelft.nl/ experiencelab/
[37] H. Alers, H. Liu, and I. Heynderickx, "Task Effects on Saliency: Comparing Attention Behavior for Free-Looking and Quality-Assessment Tasks," in preparation.
[38] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II,” http:/ / www.vqeg.org
[39] S. Winkler, “Vision Models and Quality Metrics for Image Processing Applications,” Ph.D. dissertation, Dept. Elect., EPFL, Lausanne, 2002.
[40] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, “A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms,” IEEE Transactions on Image Processing, vol. 15, no. 11, pp, 3440-3451, November 2006.
[41] D. C. Montgomery and G. C. Runger, Applied Statistics and Probability for Engineers. New York: Wiley-Interscience, 1999.
[42] J. Redi, H. Liu, P. Gastaldo, R. Zunino and I. Heynderickx, "How to Apply Spatial Saliency into Objective Metrics for JPEG Compressed Images?" in Proc. IEEE International Conference on Image Processing, pp. November 2009.
[43] U. Engelke, H. Liu, H.-J. Zepernick, I. Heynderickx and A. Maeder, “Comparing two eye-tracking databases: the effect of experimental setup and image presentation time on the creation of saliency maps,” in Proc. Picture Coding Symposium, 2010.

Chapter 7

Discussion and Conclusions

In this thesis, we have described our contributions to the development of objective image quality metrics, which are increasingly demanded due to the fast growth in nowadays digital imaging systems. Reliably predicting the extent to which humans perceive quality aspects remains an academic challenge, and current research is still far from mature. Our study specifically focused on two main research questions:

- what is the added value of adding HVS characteristics to the design of specific NR metrics?
- what is the added value of adding visual attention to the design of objective metrics?

Our findings on these two research questions are detailed below. In addition, we discuss our most significant achievements and suggestions for further research in this area.

7.1 Adding HVS Characteristics to NR Metrics

The essential component of our proposed approach towards the design of artifact specific NR metrics is the local addition of human vision characteristics to the physical features of the artifacts. Obviously, adding human vision characteristics to objective metrics of artifacts is expected to increase their reliability in predicting perceived artifact annoyance. The HVS, however, is complex and its functionality is not fully understood yet. As a result, modeling the HVS functionality to its full extent is computationally very demanding, if possible at all. Hence, from a practical point of view, especially in case of real-time implementation, adding HVS characteristics to an artifact specific NR metric needs to consider the tradeoff between accuracy and computational cost. In this thesis, we considered what aspects of the HVS to take into account and how to simulate them in a computationally efficient way.

Our investigations show that including texture and luminance masking improves the performance of the estimation of the suprathreshold visibility of blocking and ringing artifacts. Comparing the prediction reliability of a metric for perceived annoyance of blockiness or ringing with and without HVS characteristics shows an improvement of roughly 30% in the Pearson correlation coefficient, obtained by including texture and luminance masking only. At the same time, several measures are taken to maintain low complexity of the HVS-based metric. First, the metric is calculated using the luminance component of the images only (i.e. excluding chromaticity channels of the incoming signal). This simplification does not affect the performance of the metric, at least not for the artifacts considered in this thesis. Second, the HVS characteristics are only calculated at those locations in the image where the artifacts are detected. This largely reduces the computational power by avoiding modeling of the HVS in irrelevant regions. Apart from limiting the computational cost of the metric, this step also makes the quantification of artifact annoyance more reliable. Detecting blocking artifacts is relatively easy, since their

spatial location is very regular. As a consequence a relatively simple grid detection method can be used to ensure the location detection of blocking artifacts in all practical applications. Detecting the location of ringing artifacts is more difficult and largely image content dependent. The use of our proposed perceptually more meaningful edge detection method has shown its great benefit to the reliability of the ringing metric. Third, having a further reduction of computational cost in mind, masking of the HVS is implemented in a different way in our blockiness metric than in our ringing metric. For our blockiness metric, HVS masking is simply formulated as a weighting coefficient that calculates the visibility of each detected blocking artifact based on its local image content. For our ringing metric masking is applied to each line segment resulting from our edge detector, and only those regions around each line segment, in which ringing is not visually masked are extracted. In line with these findings, we can state that computational cost can be gained when including HVS aspects to a specific artifact metric by carefully adapting the design of the metric to the specific structure of the targeted artifact type.

7.2 The Added Value of Visual Attention in Objective Metrics

Researchers attempt to further improve the reliability of objective metrics by taking into account visual attention of the HVS. Modeling this aspect in an objective metric is not a trivial task, and many discussions focus on two questions: first, whether visual attention should be included in objective metrics, and second, if so, how it should be done. In our investigations, we use measured data of visual attention (i.e. eye-tracking data) in an attempt to make the results independent of the reliability of a computational attention model. As such, our investigations allow us to conclude whether and to what extent the addition of saliency can be beneficial to objective quality prediction in more general terms. From our investigations, we can conclude that if saliency is added to an objective metric, it should be the natural scene saliency (NSS) driven by the scene content and not the saliency obtained during scoring the quality of the scene. Apparently, the saliency or distraction power of the distortions present in the image is already sufficiently addressed by the metric itself. In addition, we can conclude that adding NSS improves a metric's performance in predicting image quality, but the actual amount of performance gain depends on the specific metric and on the distortion type assessed. Adding saliency seems to be less beneficial for metrics that already have a high prediction performance for a given type of distortion. Furthermore, also the image content has a strong influence on the added value of including visual attention in objective metrics. The performance gain when adding saliency seems non-existing for images without a clear region-of-interest. To include our findings in a real-time implementation of an objective metric, a computational visual attention model instead of eye-tracking data measured off-line will be needed. In that case, we expect that the actual gain in prediction accuracy is lower than what we show in this thesis, at least with the current soundness of visual attention models. Therefore, it is still far from conclusive whether yes or no visual attention should be added to an objective metric. More research here is clearly needed.

7.3 Significant Findings

In summary, this thesis has advanced the research field of NR objective quality modeling with a number of significant findings. We propose a novel design for NR metrics that quantify the perceived annoyance of a specific artifact type. The essential idea behind the approach is to precisely predict where humans perceive the specific artifacts in an image and to limit the estimation of the artifact annoyance to these regions. The advantage of this approach is that relevant HVS aspects can be explicitly simulated to improve the reliability of a metric, while the additional cost introduced by the HVS is minimized. Following this approach, a NR blockiness metric and a NR ringing metric that have the intrinsic capability of being implemented in a real-time application, are developed. The reliability of these metrics in predicting quality as perceived by observers exceeds that of alternative metrics available in literature. We believe that the proposed framework can be extended to more types of artifacts, such as wireless errors, color artifacts and temporal distortions.

Measuring and combining specific artifacts inherent in an image to determine the overall perceived quality is promising, but this approach is so far strongly limited by the insufficient progress in the design and combination of individual artifact metrics. In this thesis, we demonstrate that a NR approach based on a neural network is a simple yet efficient means for predicting overall perceived quality. Neural networks are by now well defined, and as a consequence, the issue of which features to extract to feed the neural network becomes the essential component in the metric design. Extracting a large number of features is computationally expensive and the consequent increase in the metric's complexity may affect its prediction accuracy. We have shown that using dedicated features based on artifact characteristics is highly beneficial to the reliability of the metric, while at the same time the computational effort is limited. For example, when targeting image quality assessment of a specific distortion process such as JPEG2000 compression, the use of features related to the most relevant artifact being blur enables to obtain a performance of the metric comparable to that obtained with very general (pixel-based) features, but the latter requires a larger effort for feature computation and selection. It should, however, be noted that the neural network approach only provides an approximation to the overall image quality, and it does not give any information on the actual perceived annoyance of individual artifacts occurring in the image.

7.4 Future Research

To design specific NR metrics, understanding the way human beings perceive a specific artifact type is of fundamental importance. Unfortunately, most of the subjective experiments performed by the image quality community are conducted to obtain overall quality ratings rather than artifact annoyance ratings. In other words, for the design of artifact specific NR metrics more dedicated perception experiments beyond quality scoring are needed. Collecting such data, however, is more difficult than conducting a conventional scoring experiment, basically because you ask the viewer to assess the annoyance of one specific artifact while more

artifacts that degrade the quality of the image may be present. As a consequence, training the participants becomes essential, and with that the experiment becomes more time-consuming, and so, costly. In addition, an appropriate design for perception experiments that assess artifact visibility or annoyance may largely depend on the type of artifact as well as on its specific application scenario. And so, the extension of our proposed framework on the design of a NR metric to more types of artifacts essentially requires additional subjective studies on how the targeted artifact type is perceived by the human eye. Also the datasets on overall image quality scores may be too limited for the development of NR quality metrics, especially when a machine learning approach is chosen to attack the problem. To better train and test the model, large-scale subjective tests are highly beneficial. On the other hand, it is important to be aware of the limitations of subjective data. The diversity in image content, the amount of stimuli and the number of subjects that can be included in a subjective test are usually restricted. As a consequence, the performance of objective metrics usually is tested on a limited dataset and no guarantee is given on how robust the performance of the metric is against new datasets. So, to validate and compare image quality metrics, future work should focus on collecting and distributing more reliable subjective datasets.

Readers may notice that we have not addressed some topics related to quality prediction in this thesis. One of these topics is color specific quality metrics. Our proposed metrics, and also most of the metrics available in the literature, are based on the luminance signal of an image only. The computational complexity of a metric is firmly reduced by calculating only one (i.e. the luminance) instead of all three color components. In addition, most image processing techniques also only use gray-scale images, and so, a metric based on the luminance signal only can be directly implemented in these algorithms. However, taking into account chromaticity in the development of objective metrics is worth special attention, since it is one of the essential components in visual perception. Some image distortion types specifically introduce perceived color artifacts, and color enhancement is one of the most effective ways to a normal consumer to improve the image quality. Hence, introducing chromaticity related objective metrics may be a valuable extension of the current work.

Probably more important is the extension towards video quality assessment, which in the research community on quality assessment is still in the early stages of development. Although the past years have witnessed a transition in research attention from image quality to video quality, video quality research is far from mature. To some extent video quality can be approached as a summation of the image quality of the individual frames in the video, and therefore, the development of objective metrics for image quality are an essential step towards video quality metrics. But, by only considering the quality of the individual frames, temporal aspects of artifact perception are fully neglected while they may have a significant contribution to the perceived overall quality. Unfortunately, compared to what is known about spatial aspects of human vision, our knowledge on modeling temporal aspects is very limited. We believe that extending our work to the more complex problem of video quality assessment is a promising direction, given the ubiquity of streaming videos for e.g. digital television, the internet, and digital cinema.

Publication List

Thesis

[1] **H. Liu**, "Distance Determination from Pairs of Images from Low Cost Cameras", *MSc Thesis, MSc in Signal Processing and Communications*, The University of Edinburgh, UK, August 2005.

Journals

[1] **H. Liu** and I. Heynderickx, "A Perceptually Relevant No-Reference Blockiness Metric Based on Local Image Characteristics", *EURASIP Journal on Advances in Signal Processing*, vol. 2009, 2009.

[2] **H. Liu**, N. Klomp and I. Heynderickx, "A No-Reference Metric for Perceived Ringing Artifacts in Images", *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 20, pp 529-539, April, 2010.

[3] **H. Liu**, N. Klomp and I. Heynderickx, "A Perceptually Relevant Approach to Ringing Region Detection", *IEEE Transactions on Image Processing,* vol. 19, pp. 1414-1426, June, 2010.

[4] **H. Liu** and I. Heynderickx, "Visual Attention in Objective Image Quality Assessment: based on Eye-Tracking Data", *IEEE Transactions on Circuits and Systems for Video Technology, accepted*.

[5] **H. Liu**, J. Redi, H. Alers, R. Zunino and I. Heynderickx, "An Efficient Neural-Network based No-Reference Approach to an Overall Quality Metric for JPEG and JPEG2000 Compressed Images", *submitted to Journal of Electronic Imaging*.

Conferences

[1] **H. Liu** and I. Heynderickx, "Issues in the Design of a No-Reference Metric for Perceived Blur", *IS&T/SPIE Electronic Imaging 2011, Image Quality and System Performance VIII*, January 2011.

[2] **H. Liu**, J. Wang, J. Redi, P. Le Callet and I. Heynderickx, "An Efficient No-Reference Metric for Perceived Blur", *EUVIP*, 2011.

[3] J. Redi, **H. Liu**, R. Zunino and I. Heynderickx, "Interactions of visual attention and quality perception", *IS&T/SPIE Electronic Imaging 2011, Human Vision and Electronic Imaging XVI*, vol. 7865, January 2011.

[4] U. Engelke, **H. Liu**, H.-J. Zepernick, I. Heynderickx and A. Maeder, "Comparing two eye-tracking databases: the effect of experimental setup and image presentation time on the creation of saliency maps", *Picture Coding Symposium*, 2010.

[5] **H. Liu** and I. Heynderickx, "Visual Attention Modeled with Luminance Only: from Eye-Tracking Data to Computational Models", *Fifth International Workshop on Video Processing and Quality Metrics for Consumer Electronics - VPQM 2010*, January 2010.

[6] **H. Liu**, J. Redi, H. Alers, R. Zunino and I. Heynderickx, "No-reference image quality assessment based on localized gradient statistics: application to JPEG

and JPEG2000", *IS&T/SPIE Electronic Imaging 2010, Human Vision and Electronic Imaging XV*, January 2010.

[7] H. Alers, **H. Liu**, J. Redi and I. Heynderickx, "Studying the risks of optimizing the image quality in saliency regions at the expense of background content", *IS&T/SPIE Electronic Imaging 2010, Image Quality and System Performance VII*, Jan 2010.

[8] J. Redi, **H. Liu**, H. Alers, R. Zunino and I. Heynderickx, "Comparing subjective image quality measurement methods for the creation of public databases", *IS&T/SPIE Electronic Imaging 2010, Image Quality and System Performance VII*, January 2010.

[9] **H. Liu** and I. Heynderickx, "Studying the Added Value of Visual Attention in Objective Image Quality Metrics Based on Eye Movement Data", *IEEE ICIP 2009 International Conference on Image Processing*, November 2009.

[10] J. Redi, **H. Liu**, P. Gastaldo, R. Zunino and I. Heynderickx, "How to Apply Spatial Saliency into Objective Metrics for JPEG Compressed Images?", *IEEE ICIP2009 International Conference on Image Processing*, November 2009.

[11] **H. Liu**, N. Klomp and I. Heynderickx, "A No-Reference Metric for Perceived Ringing", *VPQM-09 Fourth International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, January 2009.

[12] **H. Liu**, N. Klomp and I. Heynderickx, "Perceptually Relevant Ringing Region Detection Method", *EUSIPCO2008 The 16th European Signal Processing Conference*, August 2008.

[13] **H. Liu** and I. Heynderickx, "A No-Reference Perceptual Blockiness Metric", *IEEE ICASSP 2008 The 33rd International Conference on Acoustics, Speech, and Signal Processing*, March 2008.

[14] **H. Liu** and I. Heynderickx, "A Simplified Human Vision Model Applied to a Blocking Artifact Metric", *CAIP 2007 The 12th International Conference on Computer Analysis of Images and Patterns*, Lecture Notes in Computer Science (LNCS), Springer, August 2007.

[15] S. Kiranyaz, **H. Liu**, M. Ferreira and M. Gabbouj, "An Efficient Approach for Boundary Based Corner Detection by Maximizing Bending Ratio and Curvature", *IEEE ISSPA 2007 International Conference on Information Sciences, Signal Processing and its Applications*, February 2007.