Stereoscopic Video Quality Assessment Using Binocular Energy

Chathura Galkandage, *Student Member, IEEE,* Janko Calic, *Member, IEEE,* Safak Dogan, *Member, IEEE* and Jean-Yves Guillemaut, *Member, IEEE*

***Abstract*—Stereoscopic imaging is becoming increasingly popular. However, to ensure the best quality of experience, there is a need to develop more robust and accurate objective metrics for stereoscopic content quality assessment. Existing stereoscopic image and video metrics are either extensions of conventional 2D metrics (with added depth or disparity information) or are based on relatively simple perceptual models. Consequently, they tend to lack the accuracy and robustness required for stereoscopic content quality assessment. This paper introduces full-reference stereoscopic image and video quality metrics based on a Human Visual System (HVS) model incorporating important physiological findings on binocular vision. The proposed approach is based on the following three contributions. First, it introduces a novel HVS model extending previous models to include the phenomena of binocular suppression and recurrent excitation. Second, an image quality metric based on the novel HVS model is proposed. Finally, an optimised temporal pooling strategy is introduced to extend the metric to the video domain. Both image and video quality metrics are obtained via a training procedure to establish a relationship between subjective scores and objective measures of the HVS model. The metrics are evaluated using publicly available stereoscopic image/video databases as well as a new stereoscopic video database. An extensive experimental evaluation demonstrates the robustness of the proposed quality metrics. This indicates a considerable improvement with respect to the state-of-the-art with average correlations with subjective scores of 0.86 for the proposed stereoscopic image metric and 0.89 and 0.91 for the proposed stereoscopic video metrics.**

***Index Terms*—Stereoscopic video, HVS model, Quality metric.**

I. Introduction

STEREOSCOPIC video has become a major format for 3D multimedia. However, as a result of the inherently high bandwidth of stereoscopic content, video compression algorithms are used, which induces visual artefacts and a degradation in quality. Additionally, video delivery over mobile and wireless channels is prone to errors in transmission, further degrading the users' Quality of Experience (QoE). Due to the inefficiency and unreliability of subjective evaluations of the perceived video quality, there is a high demand for objective computational methods for estimating perceived stereoscopic video quality. There have been several proposed approaches to generate an effective and accurate quality metric for stereoscopic content, however due to their lack of generality and robustness, there is currently no widely accepted solution.

An effective method to quantify QoE must take into consideration the different stages of stereoscopic video production namely capture, processing, transmission, rendering and display. This research focuses on developing an objective measure relating to the processing stage and more specifically the compression stage. Compression of stereoscopic content typically introduces blocking and ringing artefacts both resulting in erroneous depth. Compression is therefore a major source of degradation in stereoscopic video production whose effects need to be considered to develop a reliable quality metric.

Recently, a very promising research direction has been the use of a model of human perception to devise more reliable measures of perceived quality. These approaches use a model of the Human Visual System (HVS) to mimic depth perception and in particular the binocular disparity induced by the horizontal displacement of image features between left and right views. From a physiological point of the view, the primary visual cortex of the HVS has two main types of cells namely simple cells and complex cells responsible for depth perception. Following initial processing in the receptive fields of the ganglion cells, two retinal images are passed to the primary visual cortex (V1) in the brain. The retinal information is first received by the simple cells which work in pairs corresponding to the left and right eyes. Each pair of receptive fields is connected to a complex cell where the binocular signal is generated. Recently, a quality metric based on a HVS model has been presented; the metric shows an increased performance compared to conventional quality metrics [1]. However, this type of approach has so far been limited to a relatively simple model of human perception, which does not capture accurately the behaviour of complex cells. Besides, the approach has so far remained limited to stereoscopic images.

This paper introduces a stereoscopic video quality metric based on an enhanced HVS model which for the first time is extended to the temporal domain. The approach can be broken down into three key stages each making contributions to the state-of-the-art in stereoscopic video perception and quality assessment. First, a novel HVS model is introduced to more accurately represent binocular vision. This is accomplished by incorporating additional physiological findings about human binocular vision to address the shortcomings of the previous HVS model [1]. This results in a more sophisticated rep-

C. Galkandage, J. Calic and J.-Y. Guillemaut are with the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, GU2 7XH, UK (e-mail: c.perera@surrey.ac.uk; j.calic@surrey.ac.uk; j.guillemaut@surrey.ac.uk).

S. Dogan is with the Institute for Digital Technologies, Loughborough University London, London, E15 2GZ, UK (email: s.dogan@lboro.ac.uk).

Manuscript received March 8, 2016; revised August 30, 2016 and October 12, 2016; accepted November 11, 2016.

Published in: IEEE Journal of Selected Topics in Signal Processing, 11(1):102-112, 2017. DOI: http://dx.doi.org/10.1109/JSTSP.2016.2632045

resentation capable of modelling binocular suppression and recurrent excitation - two important phenomena occurring in complex cells. This produces improved binocular signals which are more reliable indicators of depth perception. Second, a statistical analysis technique is introduced to infer a relationship between the obtained binocular signals and the subjective scores in the case of stereoscopic images. Finally, the approach is extended to the video domain via temporal pooling to infer two stereoscopic video quality metrics. The resulting stereoscopic image and video quality metrics are shown to be superior to the traditional stereoscopic metrics as well as a previous perceptual stereoscopic quality metric.

The paper is organised as follows. Section II reviews the state-of-the-art in stereoscopic quality analysis and the background on physiological models of the HVS. The proposed method is described in Section III first introducing the proposed HVS model, then describing its use to build quality metrics in both image and video domains. Experimental results are presented and discussed in Section IV; this includes a comparison against the state-of-the-art in both image and video quality metrics. Section V concludes the paper by summarising the findings and discussing avenues for future work.

II. Related Work

Finding a stereoscopic quality metric has been the subject of active research over the last decade. A key challenge over using 2D quality metrics is that the perceived depth is not easy to model objectively since it is perceived solely by the human observer. Depth information is not readily present in the typical content representations and, in addition, its estimation remains a challenging task. Instead, the left and right views must be analysed along with disparity or depth maps to acquire an indication of the perceived depth quality. This challenge has prompted a prolific research activity aimed at tackling the problem of quality assessment in stereoscopic visual media. Most of those attempts were extensions of a 2D video quality metric and as a result were not tailored specifically for stereoscopic content and lacked accuracy. In contrast, this paper focuses on the physiological process of human vision and its pathway to objective measures of stereoscopic perception.

A. Stereoscopic quality metrics based on objective analysis

There have been numerous attempts to approach the assessment of the perceived quality of stereoscopic visual content by using existing 2D quality metrics and separately analysing left and right views of stereoscopic content.

In a recent study [2], the use of established 2D quality metrics such as Peak Signal to Noise Ratio (PSNR), Structural Similarity Index Metric (SSIM), Just Noticeable Difference (JND), Picture Quality Scale (PQS) and Noise Quality Metric (NQM), was found to be less effective when extended to stereoscopic content. The evaluation compared the performance of the aforementioned metrics applied to 2D and 3D images under blur, noise and compression impairments. The observed reduction in performance with 3D images was attributed to the fact that stereoscopic perception is not only affected by image content, but also by other attributes of stereopsis such as disparity.

Disparity maps were also utilised by [3] to derive a measure of 3D perception. They concluded that the 3D content of a disparity map could not be interpreted by 2D metrics based on a fidelity score combining disparity map score and average stereoscopic score. In [4], Kaptein et al. performed subjective experiments using the same objects at different depths subject to different compression rates. They found no correlation between the depth of an object and its perceived quality. In comparison, a high correlation was found between the compression bit rates of the same object and its perceived quality. Further experiments in [5] confirmed the limitations of the usage of 2D metric for stereoscopic quality assessment.

Other approaches based on the analysis of 3D information such as disparity maps and depth maps have been proposed. For instance, in [6], Tikanmaki et al. evaluated the application of Video SSIM (VSSIM) [7] and PSNR to colour+depth sequences. They concluded the need to devise metrics specific to stereoscopic content. In another work, the application of PSNR to the binarised contours of depth maps was observed to result in no improvement in performance over PSNR applied to left and right views [8]. A good correlation to human perception was obtained when the depth maps were computed using stereoscopic images affected by low impairment using SSIM [9]. However the correlation was also found to degrade as the significance of the impairment increases.

A no-reference metric for asymmetric JPEG compression was proposed in [10]. The metric was derived from compression block pairs with and without contours. Pixel-wise difference in a block and zero crossings were used to compute the blockiness as a characteristic of 3D artefacts. In another video quality metric [11], chromatic transition and depth fluctuations were combined to predict the subjective scores.

Very recently, several attempts were made to measure the impact of compression artefacts on the perceptual quality of compressed stereoscopic video. One such attempt was the Compressed Stereoscopic Video Quality (CVSQ) metric, which considers blur, blocking artefacts and inter-view similarity [12]. The CVSQ metric relies on a disparity estimation technique and has been found not to be accurate enough due to the fact that the used features are not linearly correlated with variations in subjective perception [13].

Overall, 2D quality metrics have limited ability to predict the quality of stereoscopic content despite significant efforts to incorporate 3D information such as disparity and depth. Consequently they tend to lack the accuracy and robustness required for stereoscopic content quality assessment.

B. Stereoscopic quality metrics based on HVS modelling

Recently, quality assessment algorithms exploiting the HVS model have started to emerge. In [14], Shao et al. have proposed a perceptual full-reference quality assessment metric for stereoscopic images by considering binocular visual characteristics. The approach checks the left-right consistency and compares matching errors between corresponding pixels based on binocular disparity calculation, followed by classification

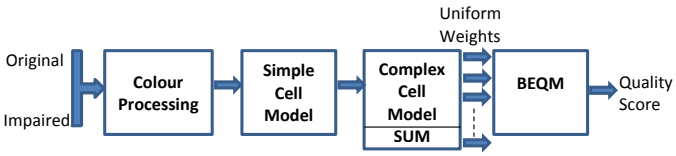


Fig. 1. System diagram for the BEQM

of the stereoscopic images into non-corresponding, binocular fusion, and binocular suppression regions. Quality assessment also considered the local phase and local amplitude maps of the original and distorted stereoscopic images as features. An overall score is obtained by integrating the binocular perception results of independently evaluated regions. The visual sensitivity for the binocular fusion and suppression regions is modelled using the Binocular Just Noticeable Difference (BJND) model.

A comprehensive set of subjective experiments was performed with stereoscopic video sequences in [13]. Sequences were encoded using both H.264/Advanced Video Coding (AVC) and High Efficiency Video Coding (HEVC) video codecs. Results of the subjective experiments with symmetrically and asymmetrically encoded stereoscopic videos were analysed using statistical techniques to identify subjective scoring patterns. As a result, a video quality metric referred to as Stereoscopic Structural Distortion (StSD) was developed. However, the metric does not consider ringing artefacts commonly present in wavelet based video codecs.

The relationship between the perceptual quality of stereoscopic images and visual information was explored in a model of binocular quality perception proposed in [15]. Based on this model, a no-reference quality metric for stereoscopic images was introduced. The proposed metric modelled the binocular quality perception of the HVS for blurring and blocking artefacts. The overall quality index considers the amount of local blur level, blocking level and visual saliency information. However, this metric was found to have a limited accuracy.

The prediction of picture quality according to human perception was investigated in [16] which conducted a systematic, comprehensive and up-to-date review of Perceptual Visual Quality Metrics (PVQMs). In this review, the signal decomposition, just-noticeable distortion, visual attention, and artefact detection were found to be aspects frequently exploited in computational modules. But, the review also highlights the need for a better understanding of the HVS mechanisms.

Based on the binocular fusion process characterising 3D human perception, a full-reference metric for quality assessment of stereoscopic images was proposed in [1]. This introduced the Binocular Energy Quality Metric (BEQM) which models the binocular signals generated by simple and complex cells. However, the ability of the binocular energy to predict human perception remains poor due to the simplicity of the complex cell model. A system diagram of BEQM is shown in Fig. 1. The proposed method introduces an improved HVS model and tailored regression methods for image and video content.

Despite the above-mentioned efforts to incorporate a model of the HVS, there is still no perceptual 3D metric capable of matching the performance of 2D quality metrics. In addition, human 3D perception has to be further explored to better understand and exploit the phenomena pertaining to binocular vision which are relevant to predict perceived visual quality.

C. HVS modelling

There have been many computational studies of stereo vision in the past, though until recently, most studies have treated depth/disparity computation mainly as a mathematical problem. Without paying close attention to physiology, one often comes up with quality metrics that work in some sense but have little to do with the mechanisms used by the brain or the HVS. There have been several attempts to address the estimation of the perceived quality of visual information by modelling the HVS. A brief introduction to the physiology of the HVS and its computational models is given here.

Images are formed on a neural tissue at the back of the eye called retina. It contains two layers holding the synaptic interconnections between the neurons and three layers of cell bodies. The images projected onto the retina are inverted, and are exhaustively pre-processed in the retina before progressing to other parts of the brain. The visual cortex that processes this information is located at the back of the brain. The primary visual cortex (V1) is the largest part of the HVS and it receives the signals from the Lateral Geniculate Nucleus (LGN) located in both hemispheres of the brain. There is a large variety of cell types in the visual cortex, responding to different kinds of stimuli, e.g. particular frequencies, colours or direction [17]. The two main types of cells in the HVS, called simple cells and complex cells, will be used in this research.

The structure of the simple cell receptive fields has been defined in [18]. The definition assumes that the role of a simple cell is dictated by the appearance of their receptive fields. Later, simple cells were identified as linear spatial filters characterised by their elongated shape composed of two antagonistic regions ON (excited) and OFF (inhibited) [19]. Further physiological experiments have showed that simple cells can be modelled using linear filters from their impulse response measured on the visual cortex. In [20], DeAngelis et al. have approximated the impulse response using a Gabor wavelet and the spatial arrangement by a two-dimensional Gabor function with ON and OFF regions corresponding to peaks and hollows of the function respectively. The above-mentioned findings have resulted in many sampling functions for simple cells which allow an image to be decomposed into perceptual channels and image elements localised in spatial and frequency domains as shown in Fig. 2. Vertical, horizontal and diagonal orientations are used in the decomposition at different spatial frequency levels, as depicted in the figure. Lower resolutions are further decomposed to a required level. In turn, this multi-resolution decomposition is used to model the perceptual channels of the simple cells.

The existence of different kinds of simple cells is an important consideration. In accordance, binocular and monocular types of cells are present with respective types of receptive fields. Monocular information from left and right retinas results in occluded information when each eye sees the world independently. The phenomenon of binocular vision of objects

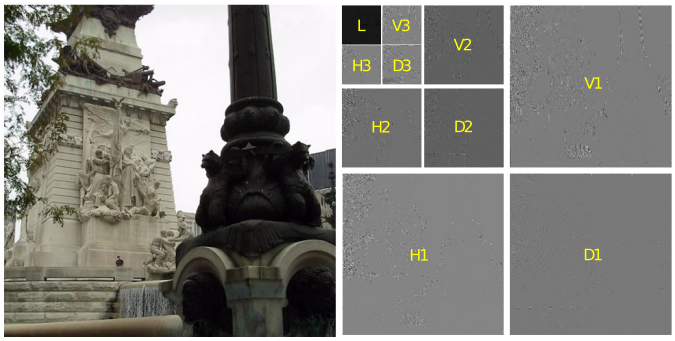


Fig. 2. Decomposition of an image into perceptual channels. Left: original image. Right: spatial frequency bands of the image. In this example, 3 orientations and 3 decomposition levels were considered, resulting in a total of 10 perceptual channels: V1, D1 and H1 correspond to the vertical, diagonal and horizontal orientations respectively at level 1 (similar notation is used for the orientations at levels 2 and 3); L is the low resolution residual

results from the existence of binocular simple cells organised in pairs and having binocular receptive fields. These binocular cells are responsible for stereoscopic perception. Several analytical models have been proposed to describe binocular simple cells. In these models, the response of a pair of binocular simple cells is often represented as a complex cell. To model the spatial frequency response based on size, amplitude, phase and orientation, directional wavelets are required. The aim is to represent the pairs of stereoscopic images using a set of complex functions.

The binocular energy is generated in the receptive fields of binocular complex cells. The spatial relationship between monocular receptive fields of complex cells and corresponding simple cells is described in [21], including the correspondence of amplitude, size, orientation and phase shift between simple and complex cells. Sensitivity to the orientation and spatial arrangement is however not inherited by complex cells from corresponding simple cells. Therefore the binocular energy generated by a complex cell depends on the shift in position and in phase between the simple cells. The field of stereoscopic quality assessment is open to further research activities to increase the reliability of 3D quality metrics to a level comparable to that of 2D quality metrics. The HVS must be modelled in an efficient way to exploit perceptual factors affecting subjective scores. Hence the previous knowledge of the HVS model will be utilised for the ultimate purpose of finding a stereoscopic video quality metric.

An improved HVS model has been proposed in our previous work [22], [23] and was used to define a stereoscopic image quality metric. This paper significantly improves upon our previous work by introducing a methodology to extend the formulation to the temporal domain and introduce a stereoscopic video quality metric. Besides, this paper performs a more comprehensive experimental evaluation using additional datasets and comparisons against the state-of-the-art.

III. Proposed Method

This section introduces full-reference stereoscopic image and video quality metrics based on an HVS model. First, a novel HVS model incorporating physiological findings unexploited in the context of quality metrics is introduced. Second, using the proposed HVS model, a stereoscopic image quality metric for compression artefacts is developed. Finally, the model is extended to the video domain via temporal pooling and a stereoscopic video quality metric for similar artefacts is constructed. All the metrics developed in this paper assume that the stereoscopic data considered is rectified (vertically registered) to ensure that corresponding features in the left and right views are vertically aligned; this requirement is easily satisfied in practice as stereoscopic content does not normally contain any significant vertical parallax.

A. Extended Binocular Energy Model (EBEM)

Similarly to [1], the proposed approach is based on modelling the behaviour of simple and complex cells. The main contribution of the proposed approach is the introduction of a more accurate model of complex cells, called the Extended Binocular Energy Model (EBEM). This new model incorporates for the first time binocular suppression and recurrent excitation, two important behaviours which were not considered in the previous Binocular Energy Model (BEM).

1) Simple cell model: The proposed approach uses a similar simple cell model to that proposed in BEM. For completeness, an overview of this model and its implementation are provided. Physiological experiments have showed that simple cells can be modelled using linear filters from their impulse response measured on the visual cortex. A Complex Wavelet Transform (CWT) is used to model the spatial frequency response of the simple cells for both luminance and chrominance components. A dual-tree method [24] is used to analyse the image using two different Discrete Wavelet Transforms (DWTs). The real and imaginary parts of the CWT are computed by applying a pair of filters, each composed of a low-pass and a high pass filter with the first couple computing the real parts of the CWT and the second couple computing the imaginary parts.

A pre-processing step is used to convert the chrominance channels in a stereoscopic image into the CIE L*a*b* [25] colour space which simulates human colour perception. This colour space uses a single channel of luminance L* and two mutually orthogonal channels of chrominance a* and b*. With the intention to represent stereoscopic images using a set of complex functions, real and imaginary parts of the response to luminance are separated using the CWT on the luminance component, whereas the chrominance response is computed using two DWTs as they are mutually orthogonal being real and imaginary parts of a complex function. Other representations such as the wavelet packet decompositions on the CWT have been proposed to increase the granularity of the directional representation [26], [27]. However, approaches purely based on the CWT have been found not to optimally characterise images containing some geometric regularity [28].

To better characterise simple cells, orthogonal bandelet bases need to be applied to the wavelet coefficients thus improving the sensitivity to the image geometry. The bandelet transform is used as it is able to adequately model this simple cells behaviour [28]. In this approach, the geometry

is computed using the best direction of the basis for each coefficient thus optimising the approximation in the presence of some geometric regularity. The bandelet construction preserves the hierarchical structure of the wavelet coefficients and utilises it to enhance any existing regularity among these coefficients and provide a more accurate image approximation. In the implementation, the set of sub-bands obtained using the analysis are organised in a quadtree of variable size following the image geometry and an orientation is computed and assigned to each block as a dyadic square depending on the coefficients. This dyadic square is characterised by its size, amplitude and orientation as a simple cell [29].

2) Complex cell model: The binocular energy is generated in the receptive fields of the binocular complex cells. The most common type of complex cells are known to perform a summation-like operation on the responses of simple cells with similar orientation preference [30]. This summation-like operation has been represented in the BEM as a binocular energy calculated using two dyadic squares representing the corresponding left and right simple cells and defined as:

$$E_{\text{SUM}}(c) = \sum_{p} \text{sum}(A_{\text{l}}^2(p,c), A_{\text{r}}^2(p,c)), \tag{1}$$

with $A_{\text{l}}(p,c)$ and $A_{\text{r}}(p,c)$ being the monocular amplitude of the binocular signals in the left and right images respectively at pixel p and for a given perceptual channel c. In this equation, the binocular energy for a given channel c is obtained by summing over the entire image pair the binocular energies contributed by each pixel p. For luminance the binocular signal is a complex function and for chrominance it is a real function. For convenience, the binocular energy scores obtained for all the considered perceptual channels can be concatenated into a vector E_{SUM}. This summation-like operation will be referred to thereafter as SUM-like for convenience. The proposed EBEM extends this model by modelling two additional characteristics of complex cells identified in physiological studies.

First, a second type of binocular energy element is added to the model in order to represent another major type of complex cells known to perform a MAX-like operation on their inputs [31]. MAX-like operation enables modelling of binocular suppression effects which occur when left and right images have undergone asymmetric impairments causing the HVS to ignore the view of one eye and perceive through the other eye. This binocular behaviour is captured by the following binocular energy term:

$$E_{\text{MAX}}(c) = \sum_{p} \max(A_{\text{l}}^2(p,c), A_{\text{r}}^2(p,c)). \tag{2}$$

Similarly to the previous case, the terms obtained for each perceptual channel c can be concatenated into a vector E_{MAX} with the same size as E_{SUM}. Hence, the complete set of binocular energy terms considering both types of operations and all perceptual channels can be represented by a single vector of binocular energies $E = [E_{\text{SUM}};E_{\text{MAX}}]$ containing twice as many elements as the number of perceptual channels.

Secondly, the model is extended to model interactions between complex cell outputs as proposed in the so-called Recurrent Excitation Model (REM) [32]. In the REM, complex cell inputs are not limited to simple cells but can also arise from other complex cells. These inputs can be thought of as secondary inputs to the complex cell model in addition to the primary simple cell inputs. These secondary inputs are modulated depending on the image content and in a deterministic manner for a given content [32]. Hence there are two important aspects of the REM that need to be considered in the implementation. First, the probability of modulation for secondary inputs needs to be determined based on the content. Second, the secondary inputs of the REM need to be representable in terms of the primary simple cell outputs. This paper introduces second order terms defined as the products of pairs of simple cell outputs to implement pairwise modulation and performs regression to learn the contributing terms and their correlation to image content.

Due to the large number of objective measures resulting from the spatial frequency analysis which includes multiple orientations and two types of complex cell characteristics representing SUM-like and MAX-like operations, the extraction of a statistical relationship between objective and subjective measures satisfying the REM presents a major challenge. This will be addressed in the following section.

B. Extended Binocular Energy Quality Metric (EBEQM)

The main idea to estimate a metric is to correlate the variations in perceptual and objective scores after normalisation with respect to a reference. This idea has been exploited in the BEQM to estimate a fidelity score Y as follows:

$$Y = aX^{\top}, \tag{3}$$

where $X = [X_1, X_2, \ldots, X_n]$ denotes the vector of normalised full-reference objective scores for a given impaired stereo image, $a = [a_1, a_2, \ldots, a_n]$ are the BEQM parameters and n is the number of coefficients in the BEM complex cell model. The vector X is obtained by normalising the vector of objective scores E of the impaired stereoscopic image defined in the previous section with respect to those of the reference (unimpaired) stereoscopic pair E_{ref} as follows:

$$X = (E_{\text{ref}} - E)/(E_{\text{ref}} + E), \tag{4}$$

where the symbol $/$ denotes an element-wise vector division. In the BEQM, the calculation of the a_i coefficients is performed under the assumption that all complex cells contribute equally, i.e. $a_i = \frac{1}{n}$ for any i, without providing any justification. This uniformity assumption, together with the simplistic BEM which only considers SUM-like complex cell operations, limits the accuracy and robustness of the BEQM.

The proposed approach addresses the previous shortcoming by estimating the contribution of each normalised objective score, thereby eliminating the uniformity assumption, and by using the more sophisticated EBEM as a basis for construction of a metric. The EBEM results in an increase in the number of objective measures due to the introduction of the second type of complex cells performing MAX-like operations. The REM can be modelled by introducing an additional vector $Z = [X_1.X_2, \ldots, X_1.X_n, X_2.X_3, \ldots, X_2.X_n, \ldots, X_{n-1}.X_n]$, whose elements consist of the product of all pairs of

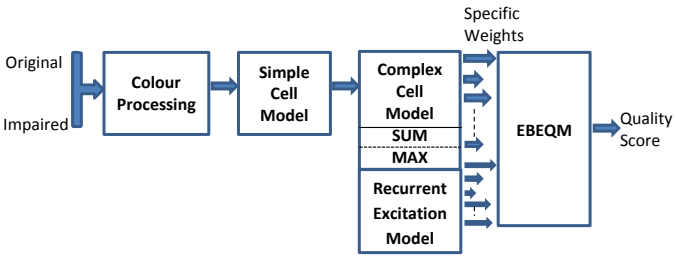


Fig. 3. System diagram for the EBEQM

normalised objective scores in X and referred to as recurrent excitation objective measures. This results in the Extended Binocular Energy Quality Metric (EBEQM) which estimates the fidelity score as:

$$Y = c + aX^{\top} + bZ^{\top}. \tag{5}$$

The EBEQM is characterised by the vector $a = [a_1, a_2, \ldots, a_n]$ of size n, the recurrent excitation coefficient vector $b = [b_{1,2}, b_{1,3}, \ldots, b_{1,n}, b_{2,3}, b_{2,4}, \ldots, b_{2,n}, \ldots, b_{n-1,n}]$ of size $n \times (n-1)/2$ and a constant c. Fig. 3 shows a block diagram of the EBEQM highlighting the additional MAX-like operation performed in complex cells, the recurrent excitation of complex cells outputs and the non-uniform coefficients for the objective scores in the metric equation.

For an N-level spatial frequency decomposition in the simple cell model, $3N + 1$ different spatial frequency sub-bands are obtained considering 3 orientations as illustrated in Fig. 2. Separate analyses are carried out on luminance (L*) and the two chrominance channels (a* and b*), resulting in $3 \times (3N + 1)$ objectives in the case of the BEM which modelled only SUM-like operations. In the EBEM, the introduction of the second type of complex cells modelling MAX-like operations increases the number of objective scores to $n = 2 \times 3 \times (3N + 1)$. In this paper, $N = 3$ was used, as in the [1], which results in a total of $n = 60$ objective scores for the proposed EBEQM (as opposed to 30 for the BEQM).

A multiple regression technique is used to map the normalised objective measures to a single subjective score. This is achieved by using stereoscopic image databases with available subjective experimental data to train a regression model. The training procedure enables estimation of a robust metric by learning the statistical relationship between the simple normalised objective scores of the binocular energy elements and the given subjective scores. The multiple regression takes the X and Y values far all the training images as input and estimates the optimal a and b vectors. The b vector in Eq. (5) is used to model the second order relationships in the REM.

Stepwise multiple linear regression was found to be the most suitable mechanism for this purpose. The learnt multiple regression models are subsequently evaluated with testing datasets to find the most sustainable relationship between the objective measures and the subjective scores. The analysis is performed using four publicly available stereoscopic image databases [14], [33], [34]. The regression model with the best testing performance, learnt from the combined databases, is defined as the metric. The details of the final metric and its calculation are described in Section IV-A.

C. Binocular Energy Video Quality Metric (BEVQM)

This paper introduces two approaches to build a stereoscopic video quality metric. First, a traditional temporal pooling approach is used to construct a video metric from the quality scores obtained for each frame using the EBEQM approach. A range of pooling strategies introduced in [35] are considered to this effect. Then an alternative approach is introduced to directly learn a video quality metric from the spatio-temporal normalised objective scores obtained from the newly proposed HVS model.

1) Temporal pooling of image quality scores: In this approach, the EBEQM is first applied to each frame of the video sequence to obtain individual estimates of the subjective score $Y(t)$ for a given frame t. Then, these estimates are combined using one of the temporal pooling strategies described in [35]. The following five temporal strategies are considered (the reader is referred to [35] for the implementation details):

a) *Mean*: this is the most widely used pooling strategy. This calculates a temporal average of the quality scores $Y(t)$.

b) *Median*: This approach requires sorting the frames based on their individual quality scores.

c) *Minkowski*: Minkowski summation for a given parameter β calculates a pooled quality metric as:

$$Y_{\text{pooled}} = \sqrt[\beta]{\frac{1}{f}\sum_{t=1}^{f} Y(t)^{\beta}}, \tag{6}$$

where f is the number of frames in the sequence considered. The mean is a special case of Minkowski summation with $\beta = 1$. In this paper a value of $\beta = 0.5$ was used as suggested in [35].

d) *Percentile*: this approach uses the bottom k percentage of frames ranked in terms of their quality score. In this paper, the 10^{th} percentile is used as suggested in [35].

e) *Gaussian weighting*: This approach applies Gaussian weighting to the ranked frame quality scores with the frames of poorer quality getting higher weights prior to summation. The optimal settings suggested in [35] are used.

2) Regression on pooled objective scores: A naïve approach to build a video metric would be to concatenate the objective scores obtained from the EBEM applied to each frame and attempt to directly perform a regression on these concatenated objective scores. This approach is problematic in several respects. First, even with a short duration video sequence, the model would have an extremely high dimensionality which would prevent reliable estimation of the model's parameters. Second, the model size would depend on the length of the video sequence, requiring some form of normalisation.

In order to remedy the above-mentioned issues, dimensionality reduction is introduced. Fig. 4 shows the binocular energy variations for some examples of objective scores. This illustrates the complexity of the problem and how minor variations over neighbouring frames and subtle changes of objective scores tend to correlate with significant changes in video content. One approach to handle temporal complexity would be to further extend the HVS model to mimic the spatio-temporal characteristics of human stereoscopic video perception. Physiological studies have identified motion sensitivity

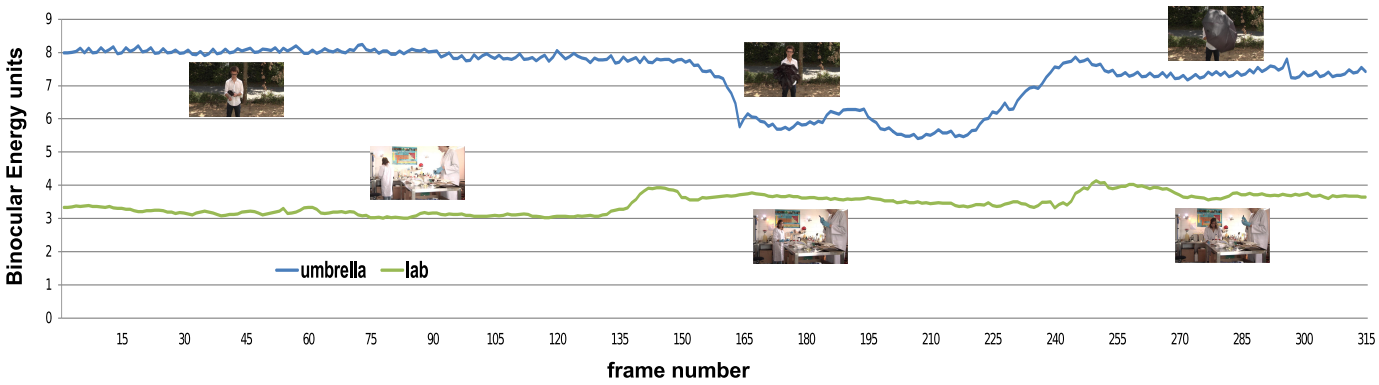


Fig. 4. Example of objective score variations for the "Umbrella" and "Lab" sequences

as an important characteristic of complex cells [1]. However, for video perception, motion sensitivity is not sufficient on its own to account for memory effects and perceptual decisions occuring in the HVS. Besides, this approach would still result in a high dimension model.

Instead, pooling of the temporal objective scores is proposed in order to extract pooled objective scores representing the temporal variations in the sequence. The method used for pooling of objective scores is referred to as adaptive temporal pooling from here onwards. In adaptive temporal pooling, only the mean and Minkowski summation methods are applicable since it would not be meaningful to rank objectives scores. In this case, the vector of characteristic objective scores representing the entire stereoscopic video sequence and obtained after Minkowski summation are defined as:

$$X_{\text{pooled}} = \sqrt[\beta]{\frac{1}{f}\sum_{t=1}^{f} X(t)^{\beta}}, \tag{7}$$

where $X(t) = [X_1(t), X_2(t), \ldots, X_n(t)]$ represents the individual objective scores at frame t calculated using the EBEM. Note that in Eq. (7) both exponentiation and root operations are performed element-wise on the vector $X(t)$. The performance of these two approaches will be discussed in the results section alongside the optimisation of the β parameter.

Having extracted pooled objectives scores, a regression analysis is carried out on the pooled objective scores in a similar fashion to the EBEQM in order to calculate a model relating pooled objective scores X_{pooled}, the corresponding recurrent excitation objective measures Z_{pooled} and the corresponding subjective score for a given stereoscopic video Y_{video}. These different parameters are related by:

$$Y_{\text{video}} = c + aX_{\text{pooled}}^{\top} + bZ_{\text{pooled}}^{\top}, \tag{8}$$

where a, b and c are the parameters describing the video metric. These retain the same dimensionality as in the previous approach. The details of the regression analysis carried out and the metric obtained, called the Binocular Energy Video Quality Metric (BEVQM), will be discussed in the next section.

IV. Results and Discussion

The outcomes of the multiple regression analysis for both image and video based objective measures are presented in this section together with the metrics obtained.

TABLE I
CHARACTERISTICS OF THE STEREOSCOPIC IMAGE DATASETS

Database name	# images	Distortion types with # quality profiles indicated in brackets	Quality profiles symmetry	# stimuli per image
LIVE 3D Image Quality - Phase II [33]	8	white-noise(9), JPEG2K(9), JPEG(9), BLUR(9), fast fading(9)	3 symmetric, 6 asymmetric	45
IVC 3D Images [34]	5	BLUR(5), JPEG(5), JPEG2K(5)	symmetric	15
Ningbo symmetric stereo [14]	12	JPEG(5), JPEG2K(5), GBLUR(5), white-noise(5), H.264(6)	symmetric	26
Ningbo asymmetric stereo [14]	10	JPEG(7), JPEG2K(10), GBLUR(10), white-noise(10)	asymmetric	37

A. EBEQM

Four different publicly available stereoscopic databases were used to build the EBEQM. All the databases consist of vertically aligned stereoscopic images. A summary of the different datasets used for training is shown in Table I.

First, an evaluation is carried out by using one database at a time to build and test models. This considers five testing configurations corresponding to each of the four individual datasets as well as an additional scenario where all datasets are merged into a single large dataset. This enables assessment of the consistency and repeatability of the regression analysis across datasets as well as its performance according to the size of the training dataset. In each testing scenario, the models are built using stimuli from a single dataset with one stimulus from the same dataset being excluded and used for testing purposes. This process is repeated for all possible training and testing subsets for each dataset in order to obtain statistically meaningful results on the performance of the models obtained. Performance is assessed by calculating the Pearson's linear correlation coefficient between the scores predicted by the metric and the subjective scores. The results obtained for each of these configurations are shown in Fig. 5. The average correlation measured using the testing set is significant being in the region of 0.8. It is observed to be consistent over all five testing scenarios. Additionally, the results obtained on the dataset combining all four datasets demonstrate the robustness of the regression model on different types of content.

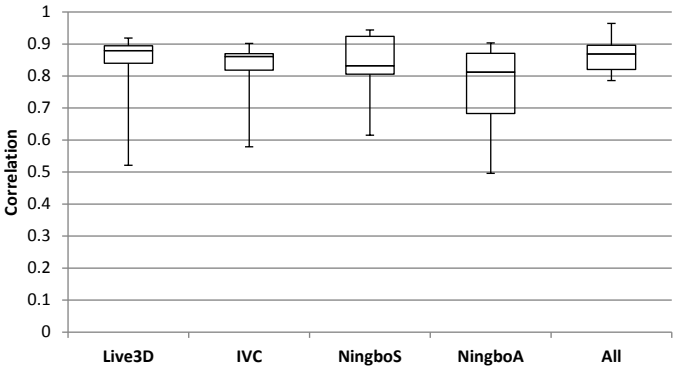


Fig. 5. Box and whisker plot showing minimum, first quartile, median, third quartile and maximum correlation for the regression models learnt and tested on individual datasets (Live3D, IVC, NingboS and NingboA) and the combined dataset (All)

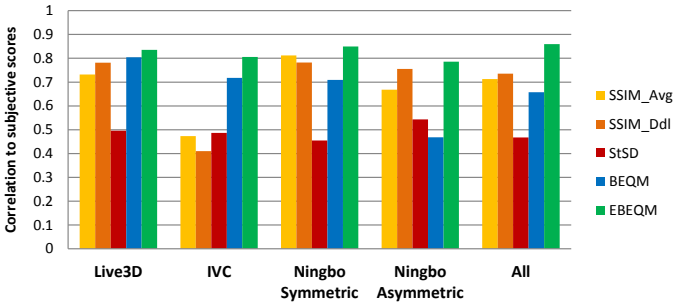


Fig. 6. Comparison of the proposed method to the state-of-the-art showing performance on individual datasets and the combined dataset

Next, the performance of the approach is compared against the following state-of-the-art image quality metrics:

a) *SSIM_avg* [7]: this is based on luminance, contrast and structural comparison.

b) *SSIM_Ddl* [34]: this uses a global 2D image distortion measure and the differences in disparity maps of stereo pairs.

c) *StSD* [13]: this computes a metric based on structural distortion, asymmetric blur and content complexity.

d) *BEQM* [1]: this uses the BEM.

For the proposed method, a model is built for each image using the other images for training and the average correlation is reported. This ensures a fair comparison with no overlap between training and testing sets. Results obtained for the different methods are shown in Fig. 6 and indicate that the proposed approach results in a significant improvement in performance across all datasets. Then, the image metric parameters a, b and c used in Eq. (5) are estimated by identifying the regression relationship of the best performing model built previously. The obtained metric consists of 122 non-zero coefficients. The most significant coefficients are listed under the EBEQM section in Table II. The complete list of non-zero coefficients is available as a supplementary spreadsheet [36]. The estimated values for the coefficients are given along with their Standard Error (SE) and pValue. The pValue represents the probability of getting the extreme results given that the null hypothesis is true whereas SE is a measure of the statistical accuracy of an estimate.

Finally, the effects of incorporating the MAX-like operation

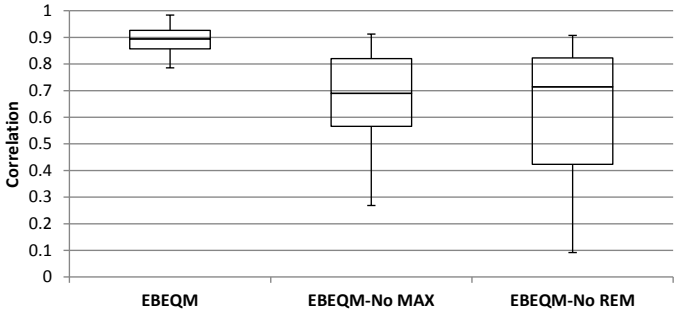


Fig. 7. Box and whisker plot showing minimum, first quartile, median, third quartile and maximum correlation for EBEQM, EBEQM without MAX-like operation and EBEQM without REM on combined dataset

and the REM into the HVS model are evaluated. To measure their separate effects, metrics are built using the combined datasets excluding either the MAX-like operation or the REM. The performance of the metrics are then compared against the full EBEQM in Fig. 7. Results indicate a significant drop in correlation when either of these physiological phenomena are excluded, demonstrating their importance in building an accurate and robust stereoscopic image quality metric.

B. BEVQM

Due to the limited availability of publicly accessible datasets with stereoscopic video subjective scores, a tailored dataset captured as part of the ROMEO project (http://www.ict-romeo.eu) is introduced and used in addition to the publicly available NAMA3DS1-CoSpaD1 dataset [37]. A summary of these datasets is given in Table III. Both databases consist of vertically aligned stereoscopic video sequences. Sample images from the sequences in the ROMEO dataset are shown in Fig. 8. In the remainder of this section, the two proposed approaches to build a stereoscopic video quality metric are discussed.

1) Temporal pooling of image quality scores: All temporal pooling methods described in Section III-C1 are applied on EBEQM as well as SSIM_avg and SSIM_Ddl estimates of each frame of the testing sequences from the dataset. The testing results for each method are shown in Fig. 9 with $\beta = 0.5$ used for the Minkowski summation method as suggested in [35]. The correlation to subjective scores is very poor in the EBEQM and even poorer in the case of the other metrics considered. The EBEQM metric employed here remained unchanged for all the frames and has therefore constant a, b and c parameters. The poor performance of this metric when applied to pooled image quality scores can be explained by the fact that subjective video scores are only defined for the entire video and do not reflect the actual scores of frames considered in isolation. This indicates that it is not possible to build an accurate and robust stereoscopic video metric by simple pooling of the frame scores and highlights the need to construct a video-specific metric. In the next section, we show how performing pooling of objective scores instead of quality scores provides additional modelling accuracy and robustness. The major advantage of performing pooling at this earlier stage is that it enables the regression to take

TABLE II
MOST SIGNIFICANT COEFFICIENTS OF EBEQM; COMPLETE LIST OF NON-ZERO COEFFICIENTS OF BEVQMβ AND BEVQMμ. COEFFICIENTS ARE LISTED IN ORDER OF DECREASING SIGNIFICANCE

EBEQM				BEVQMβ				BEVQMμ			
Coefficient	Estimate	SE	pValue	Coefficient	Estimate	SE	pValue	Coefficient	Estimate	SE	pValue
a_{19}	152.5988	12.36622	2.55E-30	c	4.771979	0.152526	5.67E-61	c	4.968098	0.135946	6.66E-68
$b_{44,50}$	-12433.3	1100.212	3.17E-26	a_3	-5.98004	0.555038	1.51E-19	a_3	-5.30282	0.594995	5.59E-15
$b_{24,50}$	19210.76	1721.203	1.09E-25	a_{32}	7.025674	0.984242	6.81E-11	a_{35}	6.663183	1.178129	1.03E-07
$b_{33,44}$	-639.45	58.73432	1.19E-24	$b_{13,32}$	-13.1242	3.40494	1.84E-04	a_{13}	-2.11905	0.40553	7.20E-07
$b_{49,56}$	-717.348	70.4097	4.59E-22	a_{41}	7.63663	2.100525	4.04E-04	$b_{13,28}$	10.92616	2.25442	3.71E-06
$b_{3,44}$	672.0649	67.6335	3.69E-21	a_{12}	-4.61159	1.364046	9.65E-04	a_{25}	-6.95057	1.474719	6.49E-06
$b_{53,56}$	248.9721	25.40696	1.14E-20	a_{11}	-7.78248	2.654097	4.00E-03	a_9	-6.94099	1.709378	8.65E-05
$b_{42,56}$	311.7277	32.50463	6.18E-20	$b_{11,32}$	9.544086	3.951371	1.72E-02	a_{58}	11.22647	2.784096	9.62E-05
a_{49}	-145.121	16.16116	7.56E-18	$b_{12,13}$	4.380798	1.844929	1.91E-02	$b_{13,35}$	-4.959	1.605626	2.49E-03
$b_{51,57}$	-322.192	36.30945	1.71E-17	a_{13}	-0.26817	0.744026	7.19E-01	a_{28}	-10.858	3.55878	2.79E-03
a_{35}	84.70863	9.554858	1.81E-17					a_{32}	2.761953	1.012062	7.28E-03
$b_{13,15}$	171.8042	20.02932	1.60E-16					a_{52}	2.494784	0.969829	1.13E-02

TABLE III
CHARACTERISTICS OF THE STEREOSCOPIC VIDEO DATASETS

Database name	# videos	Distortion types with # quality profiles indicated in brackets	Quality profiles symmetry	# stimuli per video
NAMA3DS1-CoSpaD1 [37]	10	H264(3), JPEG2K(4), downsampling(1), sharpening(1), downsampling+sharpening(1)	symmetric	10
ROMEO	6	H.264(7)	2 symmetric, 5 asymmetric	7

Fig. 8. Sample images from the ROMEO dataset

into account the temporal properties of objective scores when selecting which are the most significant to build a video metric.

2) Regression on pooled objective scores: First, experiments are conducted on the NAMA3DS1-CoSpaD1 dataset to identify the optimal β value for adaptive temporal pooling using Minkowski summation in Eq. (7). Fig. 10 shows the measured correlation as a function of the β parameter. The results indicate that performance is relatively stable with a stable peak at $\beta = 0.66$ which has been chosen as the optimal value for Minkowski summation in the remainder of the paper.

Second, the evaluation is extended to both datasets in order

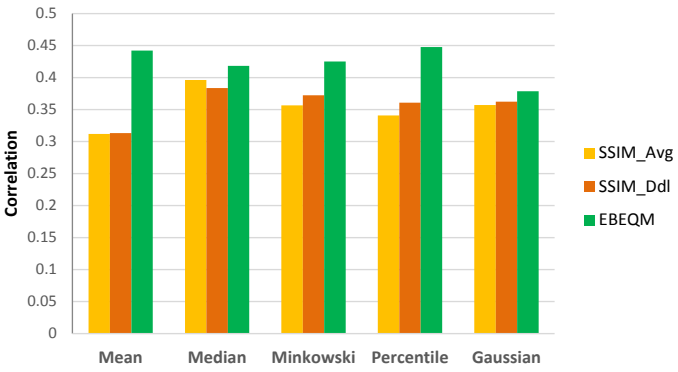


Fig. 9. Testing results for different temporal pooling techniques using stereoscopic image quality metrics

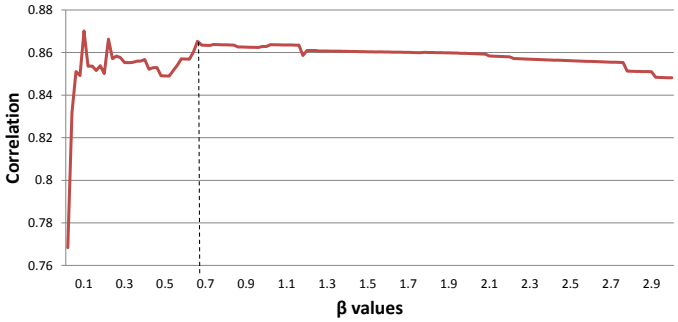


Fig. 10. Correlation as a function of β for the Minkowski summation approach

to compare the performance of the Minkowski summation method against the mean for each video, each time using all the other videos for training a model and thus ensuring that testing and training sets are disjoint. Results obtained for each video are shown in Fig. 11. All the models exhibit high correlation levels (close to 0.9 on average) with the Minkowski summation method performing marginally better consistently. Furthermore, a run-time analysis has been performed to compare the two metrics as shown in Table IV. These results indicate that an increase in run-time by a factor of 3.55 will produce an increase in correlation score from 0.89 to 0.91 when using Minkowski summation instead of the mean. While Minkowski summation is slower in terms of pure computation time, it should be noted that the run-time of pooling is typically not significant in stereoscopic video processing tasks since the cost of other operations usually dominates the run-time. Therefore, both the mean and Minkowski summation have been retained for further computations to build a metric. For convenience, the metrics developed using the mean and Minkowski summation are thereafter referred to as BEVQMμ and BEVQMβ respectively.

Finally, the performance of the proposed metrics is compared against state-of-the-art metrics on the sequences from the ROMEO dataset. This includes a comparison against the pooled EBEQM metric presented in the previous section. The results are shown in Fig. 12. The proposed metrics are the top performing methods consistently achieving correlation results in excess of 0.8 and outperforming the state-of-the-art metrics as well as the EBEQM metric. This demonstrates

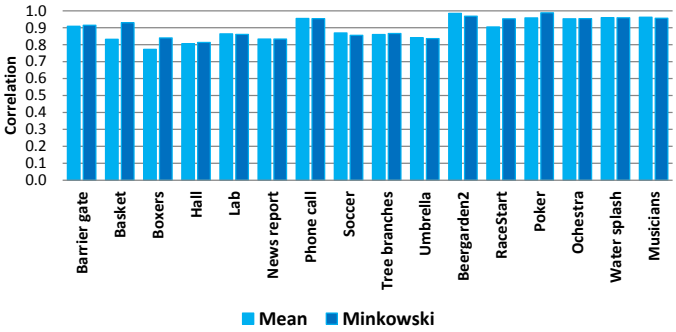


Fig. 11. Comparison of the different adaptive temporal pooling strategies for BEVQM

TABLE IV
RUN-TIME ANALYSIS OF MEAN VS MINKOWSKI SUMMATION

Pooling method	Average run time in μS	Average correlation
Mean	20	0.8918
Minkowski summation	71	0.9052

the robustness and accuracy of the proposed approach and the importance of building a video content specific quality metric. The BEVQM is computed using the best regression model in terms of testing as was the case for the EBEQM. The scores presented in Fig. 12 confirm that the BEVQMβ consistently outperforms the BEVQMμ by a small margin. These two metrics have a sparser set of coefficients than the EBEQM metric as they only contain 10 and 12 non-zero coefficients respectively. The complete list of non-zero coefficients together with their corresponding SE and pValue are provided under the BEVQMβ and BEVQMμ sections of Table II and can also be downloaded from our supplementary spreadsheet [36].

V. CONCLUSIONS AND FUTURE WORK

This paper proposed accurate and robust stereoscopic image and video quality metrics based on a novel HVS model. The first major contribution is the introduction of a novel HVS model that incorporates important physiological phenomena occurring in complex cells and so far unexploited in this context. These phenomena include a MAX-like operation, representing binocular suppression, and recurrent excitation, characterising the interactions occurring between complex cell outputs. The second contribution is a novel stereoscopic image quality metric calculated via multi-variate regression on the objective scores defined by the proposed HVS model. The third contribution is an extension to the video domain based on temporal pooling of objective scores. An extensive experimental evaluation comparing the proposed approaches against the state-of-the-art was conducted using four standard stereoscopic image datasets and two stereoscopic video datasets. Results demonstrate the robustness and accuracy of the proposed approaches to build stereoscopic quality metrics and their superiority to the state-of-the-art with average correlation to the subjective scores of 0.86 in the case of the proposed image metric and 0.89 and 0.91 in the case of the proposed video metrics. Furthermore, an experimental evaluation has validated the importance of modelling both binocular suppression and recurrent excitation in the HVS model and has shown the importance of building a video quality metric which goes beyond simple pooling of image quality scores. Future work will concentrate on extending the analysis to build a metric modelling other types of artefacts. Another avenue for future work would be to optimise the implementation for real-time applications.

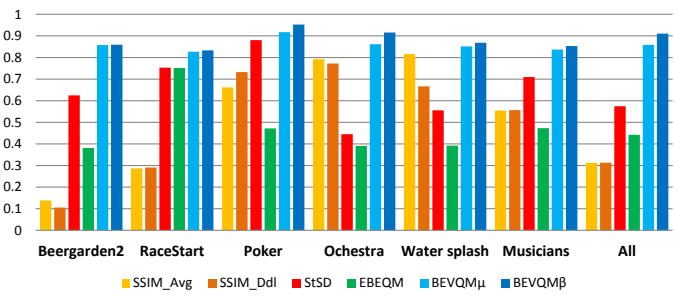


Fig. 12. Comparison of proposed methods against state-of-the-art metrics for ROMEO sequences

ACKNOWLEDGEMENTS

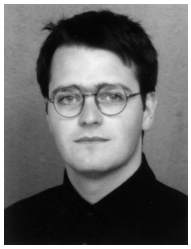
This work used stereoscopic video sequences from the ROMEO project (grant number: 287896) of the EC FP7 ICT collaborative research programme. More information about this project is available at http://www.ict-romeo.eu/.

REFERENCES

[1] R. Bensalma and M.-C. Larabi, "A perceptual metric for stereoscopic image quality assessment based on the binocular energy," *Multidimensional Systems and Signal Processing*, vol. 24, no. 2, pp. 281–316, 2013.
[2] J. You, L. Xing, A. Perkis, and X. Wang, "Perceptual quality assessment for stereoscopic images based on 2D image quality metrics and disparity analysis," in *Proc. International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2010.
[3] P. Campisi, P. Le Callet, and E. Marini, "Stereoscopic images quality assessment," in *European Conference on Signal Processing Conference*, 2007, pp. 2110–2114.
[4] R. G. Kaptein, A. Kuijsters, M. T. Lambooij, W. A. IJsselsteijn, and I. Heynderickx, "Performance evaluation of 3D-TV systems," *Electronic Imaging*, vol. 6808, 2008.
[5] L. Goldmann, F. De Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," in *IS&T/SPIE Electronic Imaging*, 2010.
[6] A. Tikanmaki, A. Gotchev, A. Smolic, and K. Miller, "Quality assessment of 3D video in rate allocation experiments," in *International Symposium on Consumer Electronics (ISCE)*, 2008.
[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.
[8] C. T. Hewage and M. G. Martini, "Reduced-reference quality metric for 3D depth map transmission," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2010.
[9] L. Xing, J. You, T. Ebrahimi, and A. Perkis, "A perceptual quality metric for stereoscopic crosstalk perception," in *International Conference on Image Processing (ICIP)*, 2010, pp. 4033–4036.
[10] R. Akhter, Z. P. Sazzad, Y. Horita, and J. Baltes, "No-reference stereoscopic image quality assessment," in *IS&T/SPIE Electronic Imaging*, vol. 7524. International Society for Optics and Photonics, 2010.
[11] D. Kim, D. Min, J. Oh, S. Jeon, and K. Sohn, "Depth map quality metric for three-dimensional video," in *IS&T/SPIE Electronic Imaging*, vol. 7237, 2009.
[12] J. Seo, X. Liu, D. Kim, and K. Sohn, "An objective video quality metric for compressed stereoscopic video," *Circuits, Systems, and Signal Processing*, vol. 31, no. 3, pp. 1089–1107, 2012.

[13] V. De Silva, H. K. Arachchi, E. Ekmekcioglu, and A. Kondoz, "Toward an impairment metric for stereoscopic video: a full-reference video quality metric to assess compressed stereoscopic video," *IEEE Transactions on Image Processing*, vol. 22, no. 9, pp. 3392–3404, 2013.
[14] F. Shao, W. Lin, S. Gu, G. Jiang, and T. Srikanthan, "Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics." *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1940–1953, 2013.
[15] S. Ryu and K. Sohn, "No-reference quality assessment for stereoscopic images based on binocular quality perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 4, pp. 591–602, 2014.
[16] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, 2011.
[17] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, and G. B. Akar, "Towards compound stereo-video quality metric: a specific encoder-based framework," in *Image Analysis and Interpretation, 2006 IEEE Southwest Symposium on*. IEEE, 2006, pp. 218–222.
[18] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.
[19] F. Campbell, G. F. Cooper, and C. Enroth-Cugell, "The spatial selectivity of the visual cells of the cat," *The Journal of Physiology*, vol. 203, no. 1, pp. 223–235, 1969.
[20] G. C. DeAngelis, I. Ohzawa, and R. D. Freeman, "Depth is encoded in the visual cortex by a specialized receptive field structure," *Nature*, vol. 352, no. 6331, pp. 156–159, 1991.
[21] Z. Liu, J. P. Gaska, L. D. Jacobson, and D. A. Pollen, "Interneuronal interaction between members of quadrature phase and anti-phase pairs in the cat's visual cortex," *Vision Research*, vol. 32, no. 7, pp. 1193–1198, 1992.
[22] G. Perera, V. De Silva, A. Kondoz, and S. Dogan, "An improved model of binocular energy calculation for full-reference stereoscopic image quality assessment," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 594–598.
[23] C. Galkandage, J. Calic, V. De Silva, and S. Dogan, "A full-reference stereoscopic image quality metric based on binocular energy and regression analysis," in *3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2015.
[24] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 123–151, 2005.
[25] J. Schanda, *Colorimetry: Understanding the CIE system*. John Wiley & Sons, 2007.
[26] J. Yang, Y. Wang, W. Xu, and Q. Dai, "Image and video denoising using adaptive dual-tree discrete wavelet packets," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 5, pp. 642–655, 2009.
[27] ——, "Image coding using dual-tree discrete wavelet transform," *IEEE Transactions on Image Processing*, vol. 17, no. 9, pp. 1555–1569, 2008.
[28] G. Peyré and S. Mallat, "Orthogonal bandelet bases for geometric images approximation," *Communications on Pure and Applied Mathematics*, vol. 61, no. 9, pp. 1173–1212, 2008.
[29] G. Peyré, "Géométrie multi-échelles pour les images et les textures," Ph.D. dissertation, Ecole Polytechnique X, 2005.
[30] I. M. Finn and D. Ferster, "Computational diversity in complex cells of cat primary visual cortex," *The Journal of Neuroscience*, vol. 27, no. 36, pp. 9638–9648, 2007.
[31] J. A. Movshon, I. Thompson, and D. Tolhurst, "Spatial and temporal contrast sensitivity of neurones in areas 17 and 18 of the cat's visual cortex." *The Journal of Physiology*, vol. 283, p. 101, 1978.
[32] L. Tao, M. Shelley, D. McLaughlin, and R. Shapley, "An egalitarian network model for the emergence of simple and complex cells in visual cortex," *Proceedings of the National Academy of Sciences*, vol. 101, no. 1, pp. 366–371, 2004.
[33] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, vol. 28, no. 9, pp. 1143–1155, 2013.
[34] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP journal on image and video processing*, vol. 2008, 2009.
[35] S. Wulf and U. Zolzer, "The impact of temporal pooling and spatio-temporal quality interaction on the prediction accuracy of video quality metrics," *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pp. 26–31, 2013.
[36] C. Galkandage, J. Calic, S. Dogan, and J.-Y. Guillemaut, "Stereoscopic video quality assessment using binocular energy (supplementary spreadsheet)," http://www.cvssp.org/projects/BinocularEnergyModels/JSTSP/, 2016.
[37] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutiérrez, and N. Garcia, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in *International Workshop on Quality of Multimedia Experience (QoMEX)*, 2012, pp. 109–114.

Chathura Galkandage is a PhD student in the Centre for Vision, Speech and Signal Processing (CVSSP) at the University of Surrey. He received the BSc (Eng.) degree (first class) in Electronics and Telecommunications Engineering from the University of Moratuwa, Sri Lanka, in 2007 and the MSc degree in Telecommunication Engineering from Middlesex University, UK, in 2010. Currently he is working on perception inspired stereoscopic visual content processing for his doctoral degree.



Janko Calic is a Visiting Lecturer in Multimedia systems and HCI at the Centre for Vision, Speech and Signal Processing, University of Surrey. His main areas of expertise are user aspects of multimedia communication systems, human-computer-interaction (HCI) and large-scale media management. He regularly reviews research in the area of multimedia signal processing and quality of experience in multimedia systems for the leading international funding bodies and publishers.

Safak Dogan is a Senior Lecturer in Multimedia Technologies at the Institute for Digital Technologies, Loughborough University London. His main areas of expertise include 2D/3D digital media processing, media adaptation and delivery, transcoding, multimedia communication systems and networks, and media quality assessments. His recent research focuses on media clouds, smart and autonomous systems. He has co-authored numerous peer-reviewed publications, and managed various EU-funded multi-national collaborative research projects.

Jean-Yves Guillemaut is a Lecturer in 3D Computer Vision at the Centre for Vision, Speech and Signal Processing, University of Surrey. His main areas of expertise include 3D reconstruction, multi-modal registration, camera calibration, free-viewpoint video and stereoscopic content production. His current research focuses on developing novel video-based modelling techniques for the reconstruction of outdoor scenes and scenes with complex surface reflectance properties. He has co-authored over 60 peer-reviewed publications.