Full-Reference Stereoscopic Video Quality Assessment Using a Motion Sensitive HVS Model

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

***Abstract*—Stereoscopic video quality assessment has become a major research topic in recent years. Existing stereoscopic video quality metrics are predominantly based on stereoscopic image quality metrics extended to the time domain via for example temporal pooling. These approaches do not explicitly consider the motion sensitivity of the Human Visual System (HVS). To address this limitation, this paper introduces a novel HVS model inspired by physiological findings characterising the motion sensitive response of complex cells in the primary visual cortex (V1 area). The proposed HVS model generalises previous HVS models, which characterised the behaviour of simple and complex cells but ignored motion sensitivity, by estimating optical flow to measure scene velocity at different scales and orientations. The local motion characteristics (direction and amplitude) are used to modulate the output of complex cells. The model is applied to develop a new type of full-reference stereoscopic video quality metrics which uniquely combine non-motion sensitive and motion sensitive energy terms to mimic the response of the HVS. A tailored two-stage multi-variate stepwise regression algorithm is introduced to determine the optimal contribution of each energy term. The two proposed stereoscopic video quality metrics are evaluated on three stereoscopic video datasets. Results indicate that they achieve average correlations with subjective scores of 0.9257 (PLCC), 0.9338 and 0.9120 (SRCC), 0.8622 and 0.8306 (KRCC), and outperform previous stereoscopic video quality metrics including other recent HVS-based metrics.**

***Index Terms*—Stereoscopic video quality assessment, human visual system, motion sensitivity, quality of experience.**

I. Introduction

STEREOSCOPIC video quality assessment remains a major research challenge due to the difficulty of devising metrics that are able to faithfully capture the complexity of human perception. Recently, there have been a number of research activities which explored how models of the Human Visual System (HVS) can be used to develop more robust stereoscopic image and video quality metrics [1]–[11]. However, these activities mostly considered stereoscopic video quality assessment as an extension of stereoscopic image quality assessment, relying on temporal pooling of stereoscopic image quality measures. A major drawback of this class of approaches is that they are unable to capture important spatio-temporal characteristics, such as the motion of objects in a scene, which require direct processing in the spatio-temporal domain. In contrast to previously reported research, this paper introduces a novel HVS model which directly encodes temporal complexity to mimic the spatio-temporal characteristics of human stereoscopic perception. The proposed HVS model generalises our previous model [8] which was also based on the HVS but excluded influence of motion sensitivity. The novel model is applied here to stereoscopic video quality assessment thereby demonstrating the importance of incorporating motion sensitivity in perceptual tasks. To the authors' knowledge, this is the first stereoscopic video quality metric based on a motion-sensitive HVS model.

Simple cells and complex cells are the main cell types in the primary visual cortex that are responsible for binocular vision in the HVS. Several physiological models have been proposed to mimic their properties [12], [13] and have been used to build image and video quality metrics [2], [8]. These models are based on the computation of binocular signals using analytical methods to estimate the perceptual quality of stereoscopic images and videos. The response to a stereoscopic input is encoded in the form of a binocular energy consisting of multiple objective scores capturing different perceptual characteristics. Physiological studies have identified motion sensitivity as an important characteristic of a significant proportion of complex cells [2]. However, to date, no model of complex cells with motion sensitivity has been developed for stereoscopic video quality assessment. This paper introduces a novel HVS model which incorporates motion sensitivity information in the computation of binocular energy, introducing new energy terms capturing motion-specific perceptual characteristics, and demonstrates its application and benefit for stereoscopic video quality assessment.

A fundamental challenge addressed in this paper relates to estimating and leveraging the level of motion present in stereoscopic videos to construct a reliable HVS model. The key insight is the introduction of a generalised complex cell model which is able to represent the behaviour of a variety of complex cells and their motion responses. This is achieved using an optical flow algorithm to extract pixel level motion information for each perceptual channel and utilising this

Manuscript received November 26, 2019; revised January 30, 2020; accepted March 1, 2020.

C. Galkandage is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK and also with the Department of Computer Science and Informatics, London South Bank University, London, SE1 0AA, UK (e-mail: c.galkandage@surrey.ac.uk).

J. Calic was with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. He is now with the BBC R&D, London, W12 7TQ, UK (e-mail: janko.calic@bbc.co.uk).

S. Dogan is with the Institute for Digital Technologies, Loughborough University London, London, E20 3BS, UK (email: s.dogan@lboro.ac.uk).

J.-Y. Guillemaut is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK (e-mail: j.guillemaut@surrey.ac.uk).

information to modulate the response of each complex cell. This results in two types of complex cells: non-motion sensitive and motion sensitive complex cells. Non-motion sensitive complex cells respond to spatial orientation regardless of whether motion is present or not, similarly to the complex cells introduced in [8]; in contrast, motion sensitive complex cells respond to spatial orientation only in the presence of motion [14], [15]. Different velocity response functions are investigated to model the behaviour of these cells as a function of the amplitude of the motion at a given orientation and scale, taking into account minimum velocity requirements.

To validate the model and demonstrate its practical use, it is applied to build a novel stereoscopic video quality metric. The metric is built by pooling both sensitive and non-motion sensitive objective scores and performing a multi-variate regression on the pooled objective scores. In the case of the motion sensitive objective scores, the level of motion in each frame is taken into account during pooling. The high dimensionality of the proposed HVS model poses computational challenges in terms of extracting a robust regression model. To address this, a tailored two-stage regression approach is proposed. In the first stage, the most significant objective scores are selected by performing a regression separately on the non-motion sensitive and the motion sensitive objective scores. In the second stage, a regression is performed on the combined set of selected non-motion sensitive and motion sensitive objective scores thereby reducing dimensionality. A comparison against state-of-the-art stereoscopic video quality metrics including the Binocular Energy Video Quality Metric (BEVQM) [8] validates the benefit of accounting for motion-sensitivity.

The novelty of the proposed method lies in accounting for both motion sensitive and non-motion sensitive complex cells of the HVS. Further, the temporal response of these complex cell types are modelled differently. In this way, modelling the HVS is expected to result in more accurate representation than by neglecting the true behaviour of the complex cells. The rest of the paper is organised as follows. Section II reviews the background on HVS modelling with a focus on motion sensitivity and the application to stereoscopic video quality assessment. Sections III and IV introduce the proposed HVS model and quality metrics. Section V evaluates the proposed approach against the state-of-the-art and discusses performance. Section VI concludes the paper by summarising the findings and discussing avenues for future research.

II. Related work

A. Physiology of the HVS

A neural tissue at the back of the eye called retina receives images. It contains two layers with synaptic interconnections between the neurons and three layers of cell bodies. The images projected onto the retina are inverted and exhaustively pre-processed before passed on 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 area) is the largest part of the HVS, which receives 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 [16].

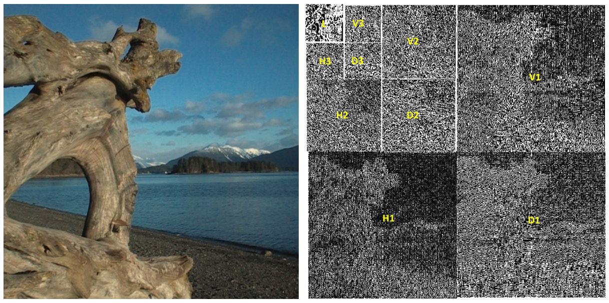


Fig. 1. Decomposition of an image into perceptual channels. Left: original image. Right: spatial-frequency bands of the image with 3 orientations and 3 decomposition levels considered and resulting in a total of 10 perceptual channels: V1, D1 and H1 correspond to the vertical, diagonal and horizontal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

Physiological experiments have shown that simple cells can be modelled using linear filters from their impulse response measured on the visual cortex. An approximation of the impulse response using a Gabor wavelet has been shown in [17], where the spatial arrangement by a two-dimensional Gabor function with ON and OFF regions correspond to peaks and hollows of the function, respectively. These 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. 1.

There is a variety of simple cells in HVS: binocular and monocular cells with their respective types of receptive fields. Monocular information from left and right retinas results in occluded information when each eye independently sees the world. Binocular vision of objects results from binocular simple cells organised in pairs with binocular receptive fields. These binocular cells are responsible for stereoscopic perception. There are several analytical models to describe simple cells and the response of a pair of binocular simple cells is often represented as a complex cell [16], [17]. The spatial-frequency response based on size, amplitude, phase and orientation can be modelled using directional wavelets with an aim 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 [18], 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. Thus, the binocular energy generated by a complex cell depends on the disparity of position and the shift in phase between the simple cells.

B. Motion sensitivity in the HVS

Direction of motion is one of the main features tested in physiological experiments. Different types of complex cells

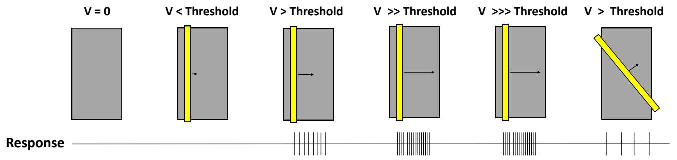


Fig. 2. Motion sensitivity of complex cells. The top part of the figure shows an oriented feature (yellow bar) moving at different velocities V in the receptive field of a complex cell, while the bottom part shows the response of the complex cell. From left to right the velocity increases from zero to a level which saturates the response of the complex cell. The right-most example shows the effect of changing the direction of motion for a given velocity.

are sensitive to different directions of motion, as illustrated in Fig. 2. This is an important phenomenon for modelling motion sensitivity of the HVS. It has been observed that there is a lower velocity threshold at which the HVS response starts and an upper velocity threshold beyond which the response saturates. Physiological studies have shown that motion sensitivity is orientation and spatial frequency selective [2]. These notable findings are at the centre of the generalised complex cell architecture proposed in this paper.

[19] showed that the velocity response of complex cells can be classified into three types: low pass, high pass and band pass responses. The first type primarily responds as a low pass filter in velocity and only a small proportion of complex cells are known to behave in this manner. The second type acts as a high pass filter in velocity and is the most common type of motion sensitive complex cells found in the HVS. It is worth noting that the cut-off velocity between these two types of filters does not occur at the same velocity threshold. The third type acts as a band pass filter and shares a maximum velocity with the second type. This type of complex cells is more common than the first type but less so than the second type. The response to velocity in all three types of complex cells can be observed to be approximately linear or uniform across the given range of operation. The study in [19] indicates that the predominant velocity response of the HVS can be modelled as a high pass filter with a linear slope or a sharp high pass filter with a uniform response after a threshold velocity. This is the model that will be implemented and evaluated in this paper.

A number of HVS models incorporating motion sensitivity and with various degrees of complexity have been proposed. A motion model based on a simple spatio-temporal concept of motion is discussed in [15]. Motion detection is formulated in terms of detecting orientation in a three-dimensional space defined by x, y, and t; the orientation exists in space-time rather than just in space. Motion in particular is filtered using appropriately oriented impulse response filters chosen as quadrature pairs sensitive to the motion direction. The combination of the outputs of two linear filters has a phase-independent-motion energy response.

If the filters' responses are squared and summed, the resulting signal gives a phase-independent measure of local motion energy within a given spatial-frequency band. The system built on these filters has motion-detecting properties with a motion response that is localised in space, time, and spatial frequency. Continuous motion, apparent motion, and motion illusions (fluted square wave and reverse phi) are basic phenomena perceived in this model. Spatio-temporal orientation can be considered as a local property of spatio-temporal stimuli and can be extracted with the same kind of simple mechanisms used for extracting spatial orientation.

A two-stage physiological model for local image velocity representation in the Middle Temporal (MT) visual area is presented in [20]. Each neuron of the MT visual area computes a weighted sum of its inputs followed by half-wave rectification, squaring, and response normalisation. Despite its simplicity, the model can account for much of the physiology of MT neurons. However, the population of model neurons is unrealistically homogeneous unlike real neurons which are irregular in comparison. Further, there is a lack of realistic temporal dynamics as the model corresponds to steady-state firing rates. This model is required to compute an estimated velocity from the responses of the MT population for the perception of speed and direction of plaid patterns.

Physiological mechanisms have been used to derive a unified model of motion and stereo vision in [21] to explain phenomena pertaining to motion-stereo interaction. In one such phenomenon, when a moving target is viewed with a neutral-density filter over one eye, it appears displaced in depth. This phenomenon is called Pulfrich's pendulum, where when the target is oscillating like a pendulum, it appears to move in an elliptical path. A demonstration of how computational modelling can help bridge the gap between physiology and perception confirms the importance of constructing computational theories of vision based on neurophysiology. However, the integrated model developed in [21] is not completely physiologically realistic.

A functional architecture of human visual motion perception is presented in [22], using four types of moving stimuli with luminance modulation, texture-contrast modulation, depth modulation and motion modulation. Seven experiments related to the four types of stimuli were conducted to determine a functional control chart. A first-order luminance system and a second-order texture-contrast system use independent motion-energy detectors, operate in parallel, and combine their outputs at an early stage. A third-order (feature-tracking) system receives inputs (features) from texture grabbers and from the lower-order motion systems. The strength of feature inputs to the third-order motion system is subject to top-down control-attention to particular features, and influences features' strengths and thereby the perceived direction of motion. The high complexity is a major concern in this architecture.

Despite the above-mentioned works, there is a large gap between physiological models of motion sensitivity and their use in quality assessment tasks. Devising novel quality metrics which incorporate the motion sensitive information available in physiological models is therefore of primary importance to achieve reliable stereoscopic video quality assessments. Sections II-C and II-D review the related stereoscopic image and video quality metrics with a focus on motion sensitivity.

C. Stereoscopic image quality metrics

In [23], established 2D quality metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM), Just Noticeable Difference (JND), Visual Information Fidelity (VIF) and Noise Quality Measure (NQM) were extended to measure the quality of stereoscopic content through averaging of the left and right view scores obtained each using the 2D quality metric. The authors observed a reduction in performance which was attributed to the fact that stereoscopic perception was not only affected by image content, but also by other attributes of stereopsis such as disparity.

Inclusion of disparity maps with 2D metrics was considered in [24] and [25]. Blur, JPEG and JPEG2000 impairments were applied symmetrically to left and right images in [24] to derive a measure of 3D perception. It was 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. Further experiments in [26] confirmed the limitations of the usage of 2D metrics for stereoscopic image quality assessments. Depth information was incorporated in different ways without directly considering the special characteristics of 3D perception such as spatial masking affected by suppression.

In [27], the performance of applying Video SSIM (VSSIM) [28] and PSNR to colour+depth sequences was evaluated. Synthesized virtual views of compressed colour and depth sequences were objectively assessed with the quality metrics. Subjective experiments showed the relative importance of colour distortions over depth distortions and the need to devise quality metrics specifically targeting stereoscopic content. In [29], a good correlation with human perception was obtained when the depth maps were computed using stereoscopic images affected by low degrees of impairment. However, the correlation was found to degrade as the significance of the impairment increases.

A full-reference metric using a product of two quality scores based on a disparity map and an extracted Cyclopean view was presented in [16]. Both quality scores were quantified using SSIM and were named as monoscopic quality (number of binocular cues preserved in images) and stereoscopic quality (via disparity map comparison). The quality metric results were then correlated to human perception. However, the results were verified using a small scale subjective experiment and did not include colour perception.

In [9], Lv et al. propose a blind (no-reference) stereoscopic image quality assessment method based on learning the receptive fields' characteristics. In this work, dictionary learning for constructing a quality lookup was used to predict subjective scores. Although good performance is reported, limited ability for dealing with asymmetrical distortion is also noted as a weakness of this approach. A full reference metric was proposed in [11] by jointly considering binocular energy and contrast perception. The approach offers mechanisms for binocular fusion and rivalry with a high prediction accuracy of perceived stereoscopic image quality.

A perceptual stereoscopic image quality approach based on modelling the properties of the primary visual cortex was proposed in [10]. This was achieved by introducing a new feature encoding approach and a tailored similarity measure that was shown to achieve high correlation with subjective scores. An effective human binocular combination model for Cyclopean image was proposed in [30], where a full-reference stereoscopic image quality assessment model was built based on binocular summation and binocular difference channels.

The survey of the state-of-the-art in stereoscopic image quality perception reveals that high accuracy prediction of subjective quality can be achieved, particularly with the recent developments in HVS-based models that outperform earlier approaches. However, better tailored approaches are required for highly reliable assessment of the stereoscopic video quality.

D. Stereoscopic video quality metrics

In [4], a comprehensive set of subjective experiments was performed with stereoscopic video sequences, which were encoded using both H.264/Advanced Video Coding (AVC) and High Efficiency Video Coding (HEVC) standards. Results of the subjective experiments on symmetrically and asymmetrically encoded stereoscopic videos were analysed using statistical techniques to reveal subjective scoring patterns. Structural distortion caused by compression was the main feature used in the metric introduced in this work. Measurement of asymmetric blur and content complexity were also used as objective measures. However, it does not consider ringing artefacts commonly present in wavelet based video codecs.

The metric presented in [31] quantifies the distortion in luminance and contrast using an approximation (variances) weighted by the mean of each pixel block to obtain the overall image distortion. The distortion on the block level is weighted to measure the frame level perceptual distortion. This metric does not account for chrominance. A stereoscopic video quality assessment method based on block-matching of left and right views via a 3D-DCT transform was proposed in [32]. However, this ignores masking effects due to motion.

In [33], spatio-temporal structural information was utilized by an algorithm which jointly represented and evaluated two views. In particular, the algorithm firstly selected salient pixels based on the results of a 3D Sobel filter. Then, the similarity of joint descriptors constructed from eigenvalues and eigenvectors of pixels in the left and right views was calculated at the pixel level. Finally, all of the local scores were pooled into one global score. This metric does not account for different degrees of the influence of salient pixels on HVS.

A novel Stereoscopic Video Quality Assessment (SVQA) metric was introduced in [34], based on the multiple visual masking characteristics of HVS, a stereoscopic just-noticeable difference model to compute the perceptual visibility for stereoscopic video. Using a stereoscopic visual attention model, stereoscopic visual saliency information was extracted first. Then, the quality maps were calculated by the similarity of the original and distorted stereoscopic videos perceptual visibility. Lack of integrity between the two models is a major drawback in this metric.

A compound stereo-video quality metric was proposed in [16] composed of monoscopic and stereoscopic quality components. Distortions causing blur, noise and contrast change were considered as monoscopic cues whereas binocular depth was the only stereoscopic cue considered. The assessment framework was based on the SSIM quality index which identified the limited perceptual measures as a major drawback.

A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video was proposed in [35]. Temporal variance, disparity variance in intra-frames, disparity variance in inter-frames and disparity distribution of frame boundary areas were used to design a no-reference stereoscopic video quality perceptual model. When the disparity in the content was high, the estimation error increased due to the incomplete disparity estimation algorithm. Direction of disparity change was not considered in this work.

[36] introduced a quality assessment model based on the observed phenomenon that spatial frequency determines view domination in the HVS. Based on the binocular fusion process characterising 3D human perception, a full-reference metric was proposed for quality assessment of stereoscopic images in [2]. The Binocular Energy Quality Metric (BEQM) introduced was modelled following a reproduction of the binocular signal generated by simple and complex cells. However, the computation of binocular energy for perceptual evaluation was poor due to the simplicity of the complex cell model. This metric was later extended to the video domain in [8] by introducing a more accurate complex cell model and an adaptive temporal pooling strategy to define the BEVQM. Despite correlating well with the subjective scores, the BEVQM lacks physiological plausibility in the time domain as it does not explicitly model motion sensitivity.

Despite considerable progress in stereoscopic video quality assessment, there is no metric making use of a motion sensitive HVS model. This paper addresses this gap by building on the recent research on HVS-based stereoscopic video quality assessment and generalising it to incorporate for the first time a physiologically inspired model of motion sensitivity. The importance of considering motion sensitivity in enhancing the accuracy of 2D video quality assessment modelling was highlighted in previous research [37], [38]. This importance is particularly magnified for realising a precise 3D video quality assessment model. As such, this paper makes its original research contribution by introducing motion sensitivity into 3D video quality assessment modelling.

III. Motion-sensitive HVS model

A. Overview of the model's architecture

The proposed motion sensitive model aims to mimic the processing taking place in the primary visual cortex (V1 area) by modelling the response of simple and complex cells. The key contribution is the introduction of a generalised complex cell architecture able to account for the behaviour of motion-sensitive complex cells as well as non-motion sensitive complex cells. The model generalises the earlier Extended Binocular Energy Model (EBEM) from [8]. A system diagram of the proposed motion sensitive model highlighting how it extends our previous model is shown in Fig. 3.

The earlier EBEM (shaded in Fig. 3) introduced a model of simple and complex cells to characterise the binocular response of the HVS. Temporal pooling of the objective scores obtained for each video frame was used to learn a metric to predict subjective perception of stereoscopic video quality. The model improved perceptual modelling compared to other approaches. However, the complex cell model lacked physiological plausibility as it ignored motion sensitivity.

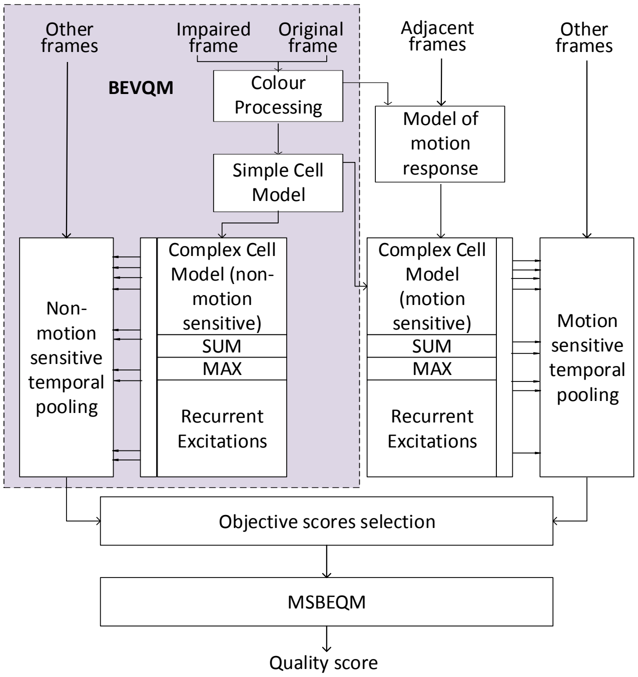


Fig. 3. System diagram of the proposed Motion Sensitive Binocular Energy Quality Metric (MSBEQM). Components from the Binocular Energy Video Quality Metric (BEVQM) system are shaded.

In contrast, the model introduced in this paper generalises the previous model by incorporating motion response maps to modulate the output of complex cells according to perceived motion. This results in a more physiologically plausible model of the response of complex cells and allows computation of a new class of objective scores capturing motion sensitive perception of stereoscopic video quality. The final model is obtained by combining motion and non-motion sensitive objective scores and is shown to result in a significant increase in its ability to predict perceived stereoscopic video quality.

Motion response maps are computed for each perceptual channel by estimating the velocity seen by the channel and then applying a velocity response function characteristic of the type of complex cell considered. Fig. 4 summarises the key processing steps to compute the motion response maps. This generalisation allows a broad variety of motion sensitive complex cell behaviours to be modelled depending on the choice of velocity response function, while retaining the ability to model simpler non-motion sensitive complex cell behaviour using a constant velocity response function. The new architecture leads to two different types of binocular energy outputs: one modelling the non-motion sensitive response of complex cells (similar to that proposed in [8]), the other one modelling the response of motion sensitive complex cells. The remainder of this section describes the key steps in the processing pipeline.

B. Simple cell model

The simple cell model used in this paper is similar to that used in [2], [8]. Stereoscopic pairs of images are represented

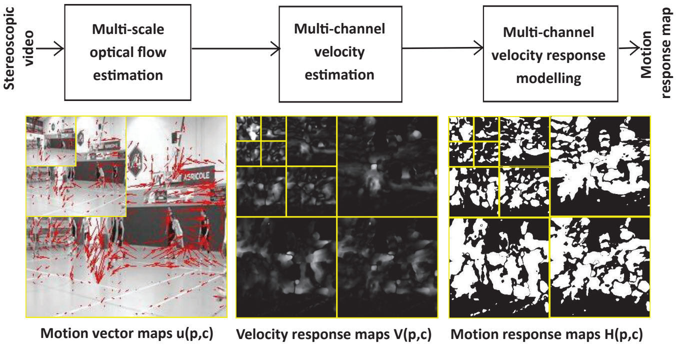

Stereoscopic video
Multi-scale optical flow estimation
Multi-channel velocity estimation
Multi-channel velocity response modelling
Motion response map
Motion vector maps u(p,c)
Velocity response maps V(p,c)
Motion response maps H(p,c)

Fig. 4. Flow chart summarising the key processing stages in the computation of the motion response maps. The output of each processing stage is shown on the bottom row under its corresponding block. The motion response map illustrated in this example was obtained using a binary velocity response function.

using complex functions $C_{\mathrm{l}}(\boldsymbol{p},c)$ and $C_{\mathrm{r}}(\boldsymbol{p},c)$ which denote the monocular signals in the left and right images at pixel $\boldsymbol{p}$ and for a given perceptual channel c. These are defined as

$$C_{\mathrm{l}}(\boldsymbol{p},c)=A_{\mathrm{l}}(\boldsymbol{p},c)e^{\phi_{\mathrm{l}}(\boldsymbol{p},c)} \text{ and } C_{\mathrm{r}}(\boldsymbol{p},c)=A_{\mathrm{r}}(\boldsymbol{p},c)e^{\phi_{\mathrm{r}}(\boldsymbol{p},c)} \tag{1}$$

where A_{l} and A_{r} denote the amplitudes while ϕ_{l} and ϕ_{r} denote the phases of the left and right signals. A brief description of the simple cell model computation is provided here, the reader being referred to [2], [8] for the full details.

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 [39] 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 filters 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 a colour space more representative of the HVS. CIE L*a*b* [40] is chosen where a single channel of luminance L* and two mutually orthogonal channels of chrominance a* and b* are used. 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. The chrominance response is computed using two DWTs as they are mutually orthogonal being real and imaginary parts of a complex function.

The bandelet transform is used to analyse the wavelet components due to its similar behaviour to simple cell characteristics [41]. The set of sub-bands obtained using the analysis are organised in a quadtree of variable size following the image geometry. An orientation is computed and assigned to each block as a dyadic square depending on the coefficients.

C. Generalised 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 SUM-like operation on the responses of simple cells with similar orientation preference [42]. Another type of complex cells are known to perform MAX-like operation [43]. Both types of operations have been modelled in [8] which defined the energy terms

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

$$E_{\mathrm{MAX}}(c) = \sum_{\boldsymbol{p}} \max(A_{\mathrm{l}}^2(\boldsymbol{p},c), A_{\mathrm{r}}^2(\boldsymbol{p},c)) \tag{3}$$

based on the functions $A_{\mathrm{l}}(\boldsymbol{p},c)$ and $A_{\mathrm{r}}(\boldsymbol{p},c)$ introduced in (1). For luminance the binocular signal is a complex function and for chrominance it is a real function. In this model, the different perceptual channels account for the different orientations and scales extracted as depicted in Fig. 1. Even though effective at modelling the orientation and scale sensitivities of the HVS, this model completely ignores motion sensitivity.

The proposed model generalises this earlier model by introducing the motion response maps $H_{\mathrm{l}}(\boldsymbol{p},c)$ and $H_{\mathrm{r}}(\boldsymbol{p},c)$ characterising the motion response at a pixel $\boldsymbol{p}$ for a given perceptual channel c in the left and right images respectively. Hence, the following energy terms are defined:

$$E_{\mathrm{SUM}}(c) = \sum_{\boldsymbol{p}} \mathrm{sum}(H_{\mathrm{l}}(\boldsymbol{p},c)A_{\mathrm{l}}^2(\boldsymbol{p},c), H_{\mathrm{r}}(\boldsymbol{p},c)A_{\mathrm{r}}^2(\boldsymbol{p},c)) \tag{4}$$

$$E_{\mathrm{MAX}}(c) = \sum_{\boldsymbol{p}} \max(H_{\mathrm{l}}(\boldsymbol{p},c)A_{\mathrm{l}}^2(\boldsymbol{p},c), H_{\mathrm{r}}(\boldsymbol{p},c)A_{\mathrm{r}}^2(\boldsymbol{p},c)) \tag{5}$$

In these equations, the binocular energy for a given channel c is obtained by first weighting the amplitude of the monocular signals in the left and right images using their respective motion response maps at each pixel, and then summing the resulting binocular energies contributed by each pixel $\boldsymbol{p}$ over the entire image. This allows complex cells to respond selectively to a particular motion.

The motion response maps for the left and right images are defined respectively as

$$H_{\mathrm{l}}(\boldsymbol{p},c) = h(V_{\mathrm{l}}(\boldsymbol{p},c)) \quad \text{and} \quad H_{\mathrm{r}}(\boldsymbol{p},c) = h(V_{\mathrm{r}}(\boldsymbol{p},c)) \tag{6}$$

where $V_{\mathrm{l}}(\boldsymbol{p},c)$ and $V_{\mathrm{r}}(\boldsymbol{p},c)$ denote the velocity maps in the left and right images, and h is the velocity response function of the type of complex cell considered.

The velocity maps $V_{\mathrm{l}}(\boldsymbol{p},c)$ and $V_{\mathrm{r}}(\boldsymbol{p},c)$ represent the amplitude of the motion at pixel $\boldsymbol{p}$ for a given perceptual channel c in the left and right views respectively. This requires a dense estimate of scene motion characterising the displacement at each pixel in the pair of image frames. It should be noted that the amplitude of the motion at a given pixel is dependent on the perceptual channel considered since it depends on both scale and orientation. Velocity map estimation will be discussed in more detail in Section III-D.

The velocity response function h is specific to a given type of complex cell. In the case of a non-motion sensitive complex cell, this is a constant function. In the case of a motion sensitive complex cell, this is a motion dependent function with profile depending on the nature of the complex cell. The motion model considered in this paper is based on implementing a high-pass filter behaviour since previous research has identified this type of behaviour as predominant [19]. The definition of the velocity response function and its effect will be discussed in more detail in Section III-E.

A two layer architecture containing both motion-sensitive (characterised by a velocity response function h_{motion}) and non-motion sensitive (characterised by a constant velocity response function h_{still}) complex cell models is considered in this paper. The binocular energy scores obtained for the different SUM and MAX operations and the different perceptual channels can be concatenated into vectors $\boldsymbol{E}_{\text{still}}$ and $\boldsymbol{E}_{\text{motion}}$ in the case of the non-motion sensitive and motion-sensitive complex cell models. To the authors' knowledge, there is no physiological evidence to suggest what proportion of complex cells response is related to motion. Hence both motion sensitive and non-motion sensitive models are considered in equal proportion and the contribution of different types of complex cells will be learnt later on together with the specific weights of each objective scores when building a metric. Hence, this results in a vector of binocular energy scores with four times as many elements as the number of perceptual channels considered (half of the binocular energy term relating to motion sensitive complex cells, the other half being non-motion sensitive).

Similarly to [8], the proposed approach models the interactions between complex cell outputs using a Recurrent Excitation Model (REM) where the output of one complex cell is modulated by the output of another complex cell according to the physiological findings reported in [44]. In the proposed model, the two layers do not converge until a common REM combines them using a regression model to produce final binocular energy elements. This generalises the previous approach by allowing modulation across complex cells with different types of motion response as well as complex cells with the same motion response. The remaining of this section provides more detail on the velocity map estimation and the definition of the velocity response function.

D. Velocity map estimation

A per pixel measure of velocity for each perceptual channel c in both left and right images is required in order to weigh the contribution of each pixel when computing the binocular energy scores in (4) and (5) and thereby represent the orientation selectivity of motion sensitive complex cells. A two-stage approach is proposed to efficiently compute the velocity maps.

a) Multi-scale optical flow estimation: First, an optical flow algorithm is used to estimate the left and right motion vectors $\boldsymbol{u}_{\text{l}}(\boldsymbol{p},c)$ and $\boldsymbol{u}_{\text{r}}(\boldsymbol{p},c)$ at each pixel $\boldsymbol{p}$ and for each perceptual channel c. The dense optical flow algorithm proposed by Farnebäck [45] is used in this paper using both the previous and the next frame to estimate the motion vectors at any given frame. The algorithm was chosen for its computational efficiency and its robustness at the time our study was performed. Optical flow is calculated separately for the left and right views. The perceived optical flow is dependent on the scale considered. For example, optical flow induced by the motion of a high frequency texture may only be visible at high resolution, disappearing when the texture becomes blurred at the lower levels of resolution. Similarly, small scale motion may only be perceptible at the higher resolution levels as its amplitude may be too small to generate a response at the lower resolution levels. Hence a multi-scale approach is used to compute the optical flow at each scale (three scales corresponding to three decomposition levels are considered in this paper). To reduce computational complexity and improve the accuracy of the motion vectors, optical flow is computed on the luminance channel only. Therefore, each image requires only three optical flow computations at the different scales considered. Optical flow estimation may be prone to inaccuracies in the presence of rapid scene motion. To some extent, the proposed HVS architecture is resilient to such errors as it does not require a very precise estimate of motion as long as the algorithm is able to distinguish pixels associated with moving scene points from static scene points, especially when using a binary velocity response function as discussed in Section III-E. Also, summation over the image provides robustness by effectively weighting down the contribution of outlier pixels with inaccurate flow.

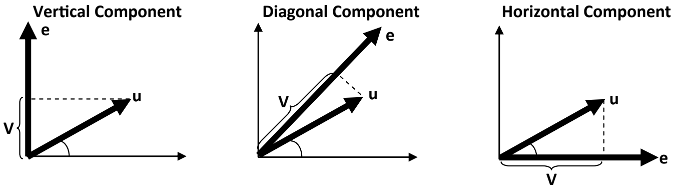


Fig. 5. Illustration of the decomposition of the motion vectors $\boldsymbol{u}_{\text{l}}(\boldsymbol{p},c)$ and $\boldsymbol{u}_{\text{r}}(\boldsymbol{p},c)$ into channel-dependent velocity components $V_{\text{l}}(\boldsymbol{p},c)$ and $V_{\text{r}}(\boldsymbol{p},c)$. For clarity, all indices have been omitted in the figure. One example is provided for each orientation in the decomposition.

b) Multi-channel velocity estimation: Second, the amount of motion in the left and right images at each pixel $\boldsymbol{p}$ for a given perceptual channel c is calculated in order to define the velocity maps. These are both scale and orientation dependent. The multi-channel image decomposition used in this paper considers three different orientations (horizontal, vertical and diagonal) at three scales and a low resolution residual. Denoting by $\boldsymbol{e}_c$ the unit vector corresponding to the orientation and scale used in the perceptual channel c, the left and right velocity components at pixel $\boldsymbol{p}$ in channel c are given by

$$V_{\text{l}}(\boldsymbol{p},c) = |\boldsymbol{u}_{\text{l}}(\boldsymbol{p},c)\cdot\boldsymbol{e}_c| \quad \text{and} \quad V_{\text{r}}(\boldsymbol{p},c) = |\boldsymbol{u}_{\text{r}}(\boldsymbol{p},c)\cdot\boldsymbol{e}_c| \tag{7}$$

in the case of the perceptual channels representing scale and orientation. This is illustrated in Fig. 5. For the last channel representing the low resolution residual, the velocity components in the left and right images are given by

$$V_{\text{l}}(\boldsymbol{p},c) = \|\boldsymbol{u}_{\text{l}}(\boldsymbol{p},c)\| \quad \text{and} \quad V_{\text{r}}(\boldsymbol{p},c) = \|\boldsymbol{u}_{\text{r}}(\boldsymbol{p},c)\| \tag{8}$$

The unit vectors $\boldsymbol{e}_c$, with respect to which motion is measured, define the three orientations (horizontal, vertical and diagonal) and the three scales with scale halved when moving from one level to the next. A total of only nine projections and one magnitude are required to compute all the velocity components for a given image. This avoids separate computation of the velocity components for the luminance and chrominance channels which share the same velocity maps.

E. Velocity response function

The velocity response of a given type of complex cell is represented using the velocity response function h. The proposed formulation is generic and versatile in that it is able to model the behaviour of both non-motion sensitive and motion sensitive complex cells including the various types of responses discussed in the literature. This paper considers the most common type of motion sensitive complex cells which are known to behave as high pass filters. The exact profile of the velocity response function for this type of complex cell is unclear. Hence two different models are considered: a binary velocity response model and a linear velocity response model as illustrated in Fig. 6.

The binary velocity response model uses a binary velocity response function h_{bin} to reject pixels with perceived velocity falling below a given threshold V_{bin} and accepts all other pixels with equal weight. This threshold is determined empirically as discussed in Section V-A. It is defined as follows:

$$h_{\text{bin}}(V) = \begin{cases} 0 & \text{if} \quad V < V_{\text{bin}} \\ 1 & \text{if} \quad V \geqslant V_{\text{bin}} \end{cases} \tag{9}$$

In contrast, the linear velocity response model uses a binary velocity response function h_{lin} to reject pixels with small motion while weighting linearly the contribution of pixels exceeding the minimum threshold V_{lin}. It is defined as follows:

$$h_{\text{lin}}(V) = \begin{cases} 0 & \text{if} \quad V < V_{\text{lin}} \\ V & \text{if} \quad V \geqslant V_{\text{lin}} \end{cases} \tag{10}$$

The binary velocity response function is a simple yet effective way of distinguishing moving scene points from static ones with the merit of being resilient to inaccuracies in optical flow estimation since it does not consider the exact amplitude of a given velocity component as long as it exceeds the minimum threshold. The linear velocity response function provides a finer grain analysis by offering the ability to take into account the actual motion amplitude when computing the binocular energy but may be more sensitive to inaccuracies in optical flow estimation. Both models will be investigated to build the stereoscopic video quality metrics in Section V.

Both models require the use of a minimum motion threshold which must be appropriately set. The threshold needs to be larger than the noise level present in the input video and smaller than the level at which object motion becomes noticeable. This threshold must also be able to mitigate any noise measured in velocity and to represent the motion sensitivity's lower threshold of the velocity response. In this paper, a common threshold of 3 pixels, measured at the resolution of the first decomposition level (highest resolution image), was used for both models. The same threshold is used at all decomposition levels since scale changes are accounted by appropriate changes in the unit vectors $\boldsymbol{e}_c$ with respect to which motion vectors are expressed. Further, to address the variability in input video resolution, all input frames are normalised to a 512×512 pixel resolution with referenced to which the threshold is defined. Further discussion and analysis of the effect of the threshold and justification of the choice of value is provided in Section V-A.

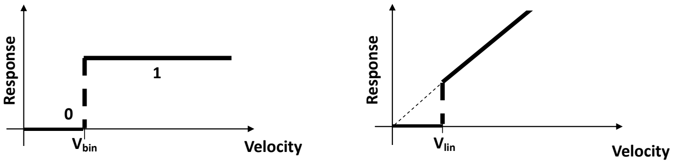


Fig. 6. Velocity response model. Left: Binary velocity response function. Right: Linear velocity response function.

IV. Motion Sensitive Binocular Energy Quality Metric (MSBEQM)

This section introduces two full reference motion-sensitive stereoscopic video quality metrics based on the generalised HVS model introduced in the previous section.

A. Normalised motion-sensitive objective scores

Let us consider a given frame t in a video sequence. The generalised HVS model can be used to calculate two sets of objective scores: $\boldsymbol{E}_{\text{still}}(t)$, representing the behaviour of non-motion sensitive complex cells and obtained using a constant velocity response function $h_{\text{still}}(V) = 1$ for any V, and $\boldsymbol{E}_{\text{motion}}(t)$, representing the behaviour of motion sensitive complex cells and obtained using velocity response function h_{motion} (two types of velocity response functions h_{bin} and h_{lin} are considered). These can be concatenated into a single vector containing all energy terms $\boldsymbol{E}(t) = [\boldsymbol{E}_{\text{still}}(t); \boldsymbol{E}_{\text{motion}}(t)]$. Similarly, the non-motion sensitive and the motion sensitive objective scores for the same frame reference stereoscopic pair can be calculated and concatenated into a vector $\boldsymbol{E}_{\text{ref}}(t)$ with the same dimension as $\boldsymbol{E}(t)$. $\boldsymbol{E}_{\text{ref}}(t)$ and $\boldsymbol{E}(t)$ are then combined to compute a vector $\boldsymbol{X}(t) = [\boldsymbol{X}_{\text{still}}(t), \boldsymbol{X}_{\text{motion}}(t)] = [X_1(t), X_2(t), \ldots, X_n(t)]$ of normalised objective scores for the given frame and defined by:

$$\boldsymbol{X}(t) = (\boldsymbol{E}_{\text{ref}}(t) - \boldsymbol{E}(t))/(\boldsymbol{E}_{\text{ref}}(t) + \boldsymbol{E}(t)) \tag{11}$$

where the symbol / denotes an element-wise vector division.

Further, a vector $\boldsymbol{Z}(t) = [X_1(t).X_2(t), \ldots, X_1(t).X_n(t), X_2(t).X_3(t), \ldots, X_2(t).X_n(t), \ldots, X_{n-1}(t).X_n(t)]$ is also introduced. It consists of the product of all pairs of normalised objective scores in $\boldsymbol{X}(t)$ and is used to implement the REM which allows complex cells to mutually interact and modulate their outputs. Unlike the previous work in which this effect was limited to non-motion sensitive complex cells, this allows different types of complex cells to modulate their outputs.

For an N-level spatial frequency decomposition in the simple cell model, $3N+1$ different spatial frequency subbands are obtained considering 3 orientations as illustrated in Fig. 1. Separate analyses are carried out on the luminance (L*) and the two chrominance channels (a* and b*), resulting in $2 \times 3 \times (3N+1)$ objective scores (half of them representing SUM-like operations, while the other half modelling MAX-like operations). The proposed model considers both motion sensitive and non-motion sensitive objective scores thus doubling the number of objective scores and resulting in a total of $n = 12 \times (3N+1)$. In this paper, $N = 3$ was used, as in [2], [8], which results in a total of $n = 120$ objective scores for the proposed MSBEQM (as opposed to 30 for the BEQM and 60 for the EBEQM) for each frame. The increased dimensionality when considering each frame poses a major challenge in terms of extracting a reliable model via regression. The remainder of this section introduces a robust approach which makes use of dimensionality reduction to build a metric.

B. Motion response weighted temporal pooling

The objective scores for the different frames are combined via temporal pooling based on Minkowski summation which was found to be the most effective in [8] in the case of non-motion sensitive objective scores. Unlike the approach proposed in [8], which equally weights the contribution of all frames, a motion response weighted temporal pooling approach is proposed here. The key idea is to measure the extent to which a particular type of motion is represented within each frame and use this information to inform the choice of pooling weight. This measure of motion at a given frame is referred to hereafter as the motion support. More formally, the motion support at frame t for a given perceptual channel c is defined as the number of pixels with non-zero motion response, that is the number of pixels $\boldsymbol{p}$ such that $h(V_l(\boldsymbol{p}, c)) > 0$ in the case of the left view (and similarly in the case of the right view). Denoting the motion support by $w_h(t)$, the sequence of values taken by the i^{th} objective score over time $X_i(t)$ with $t = 1, \dots, f$ are pooled into a single motion-response weighted objective score

$$\bar{X}_i = \sqrt[\beta]{\frac{1}{f}\sum_{t=1}^{f} |X_i(t)w_h(t)|^{\beta}} \tag{12}$$

Here, β is the Minkowski parameter set to 0.66. This value was found to be optimal in the case of the BEVQM and was observed to work well in the generalised model proposed here.

This generalises the previous approach by allowing to take into account the motion support at each frame for a particular type of motion response. It should be noted that in the case of a non-motion sensitive complex cell, the generalised approach is equivalent to the traditional temporal pooling approach since the number of pixels with non-zero motion response is constant and equal to the total number of pixels.

C. Motion sensitive stereoscopic video quality metric

Having computed the normalised objective scores and pooled them into objective scores reflecting both non-motion and motion sensitive complex cell behaviours, the next step is to identify a relationship expressing the subjective score Y as a function of the pooled objective scores $\bar{\boldsymbol{X}} = [\bar{\boldsymbol{X}}_{\text{still}}, \bar{\boldsymbol{X}}_{\text{motion}}] = [\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n]$ and their recurrent variables $\bar{\boldsymbol{Z}} = [\bar{X}_1.\bar{X}_2, \dots, \bar{X}_1.\bar{X}_n, \bar{X}_2.\bar{X}_3, \dots, \bar{X}_2.\bar{X}_n, \dots, \bar{X}_{n-1}.\bar{X}_n]$. In [8], only one type of objective scores was used. However, in this paper, there are two types of objective scores representing motion sensitivity and non-motion sensitivity. The resulting increase in the number of objective scores (120 in total) poses a challenge in identifying a metric as this considerably increases run-time but also affects the convergence of the regression technique. To address this challenge, a two-stage approach is proposed where regression is first performed separately on the non-motion sensitive and the motion sensitive objective scores to select the meaningful objective scores for each type of the motion response. In the second stage, regression is performed on the reduced set of objective scores which have been selected in the first stage. An overview of the approach is given in Fig. 3.

1) Initial regression and objective scores selection: First, two separate multi-variate regressions are performed to determine the relationships predicting the subjective score from the non-motion and the motion sensitive coefficients respectively:

$$Y = k_{\text{still}} + \boldsymbol{a}_{\text{still}}\bar{\boldsymbol{X}}_{\text{still}}^{\top} + \boldsymbol{b}_{\text{still}}\bar{\boldsymbol{Z}}_{\text{still}}^{\top} \tag{13}$$

$$Y = k_{\text{motion}} + \boldsymbol{a}_{\text{motion}}\bar{\boldsymbol{X}}_{\text{motion}}^{\top} + \boldsymbol{b}_{\text{motion}}\bar{\boldsymbol{Z}}_{\text{motion}}^{\top} \tag{14}$$

In these Equations, the vectors $\boldsymbol{a}_{\text{still}}$, $\boldsymbol{b}_{\text{still}}$, $\boldsymbol{a}_{\text{motion}}$ and $\boldsymbol{b}_{\text{motion}}$ and the constants k_{still} and k_{motion} denote the regression coefficients for the two models. As the HVS model requires cross relationships among objective outputs to meet the recurrent excitation in the complex cells model, the recurrent objective scores $\bar{\boldsymbol{Z}}_{\text{still}} = [\bar{X}_1.\bar{X}_2, \dots, \bar{X}_1.\bar{X}_{\frac{n}{2}}, \bar{X}_2.\bar{X}_3, \dots, \bar{X}_2.\bar{X}_{\frac{n}{2}}, \dots, \bar{X}_{\frac{n}{2}-1}.\bar{X}_{\frac{n}{2}}]$ and $\bar{\boldsymbol{Z}}_{\text{motion}} = [\bar{X}_{\frac{n}{2}+1}.\bar{X}_{\frac{n}{2}+2}, \dots, \bar{X}_{\frac{n}{2}+1}.\bar{X}_n, \bar{X}_{\frac{n}{2}+2}.\bar{X}_{\frac{n}{2}+3}, \dots, \bar{X}_{\frac{n}{2}+2}.\bar{X}_n, \dots, \bar{X}_{n-1}.\bar{X}_n]$ are required.

Due to the number of components in the analysis, a suppression technique is required to remove terms which are not required to stabilise the regression model. Therefore, stepwise linear regression is used over linear regression due to its ability to suppress the least meaningful components from the analysis [46]. The stepwise regression results in models containing a relatively small number of non-zero coefficients (approximately 20 overall). Only these selected objective scores will be considered in the final regression stage thus significantly reducing the pool of objective scores used in the final regression.

2) Final regression on selected objective scores: The final regression is performed by considering the selected objective scores for each type of motion. These are typically of significantly smaller size than the complete set of objective scores. The final multi-variate stepwise regression is performed to estimate the following relationship

$$Y = k + \boldsymbol{a}\bar{\boldsymbol{X}}^{\top} + \boldsymbol{b}\bar{\boldsymbol{Z}}^{\top} \tag{15}$$

where the constant k and the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are the regression parameters defining the metric. All values in $\boldsymbol{a}$ and $\boldsymbol{b}$ corresponding to non-selected objective scores are enforced to be zero. The two-stage approach results in a metric with a significantly reduced number of non-zero coefficients and a reduction in computation time by several orders of magnitude compared to a classical single-stage approach which does not perform objective score selection. Objective score selection and the obtained metrics will be discussed in Section V.

V. Results and discussion

The proposed motion sensitive approach is evaluated using the ROMEO project[1] dataset [4] and the publicly available

[1] https://cordis.europa.eu/project/rcn/100106_en.html

TABLE I
CHARACTERISTICS OF THE STEREOSCOPIC VIDEO DATASETS

dataset name	Number of SRC	Distortion types	Number of HRC	Resolution (pixels)
NAMA3DS1-CoSpaD1 [47]	10	H264, JPEG2K, downsample, sharpen	10	1920×1080
ROMEO [4]	6	H.264	7	960×1080
Waterloo 3D-VQA [48]	6	H.264	44	1024×768

datasets NAMA3DS1-CoSpaD1 [47] and Waterloo 3D-VQA [48]. All of the datasets consist of stereoscopic video sequences with additional information on associated subjective scores. The characteristics of these datasets are summarised in Table I and sample images from the different sequences from each dataset are shown in our supplementary report [49].

Two separate evaluations are conducted. First, the effects of the different approaches to model motion sensitivity and perform regression are evaluated and discussed. The analysis is performed in a leave-one-out fashion on the combined NAMA3DS1-CoSpaD1 and ROMEO datasets. Then, two motion sensitive binocular energy video quality metrics are built and evaluated against existing metrics to validate the importance of accounting for motion sensitivity. This is the core part of the evaluation which is performed by training and validating models on the NAMA3DS1-CoSpaD1 dataset and then evaluating the models independently on the ROMEO and Waterloo 3D-VQA datasets.

A. Evaluation of the effects of motion sensitivity

We proposed different approaches to model motion sensitivity depending on the motion model used and the regression approach performed on the combined set of objective scores. This results in four possible combinations of methods. To better understand the effects of the different objective scores, the models built from purely non-motion sensitive and purely motion-sensitive objective scores are also evaluated. The following seven models are therefore considered:

- **NoMo:** Regression on only the non-motion sensitive objective scores (similar to the EBEQM),
- **MoBin:** Regression on only the motion sensitive objective scores with binary velocity response function,
- **MoLin:** Same as above with linear velocity response function,
- **ComBin:** Single-stage regression on combined non-motion sensitive and the motion sensitive objective scores with binary velocity response function,
- **ComLin:** Same as above with linear velocity response function,
- **SelBin:** Two-stage regression on combined non-motion sensitive and the motion sensitive objective scores with binary velocity response function,
- **SelLin:** Same as above with linear velocity response function.

All approaches are evaluated on the combined NAMA3DS1-CoSpaD1 and ROMEO datasets in a leave-one-out fashion where each sequence is excluded in turn and used for testing purposes while the other sequences are used for training.

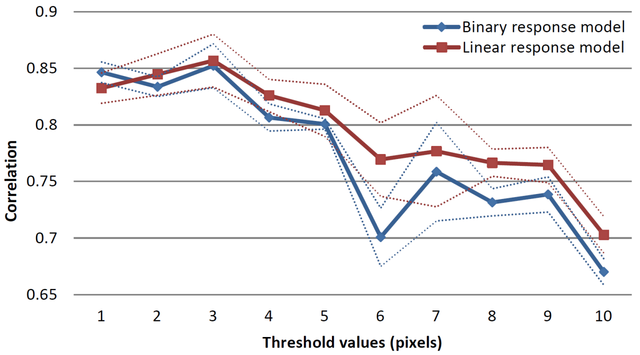


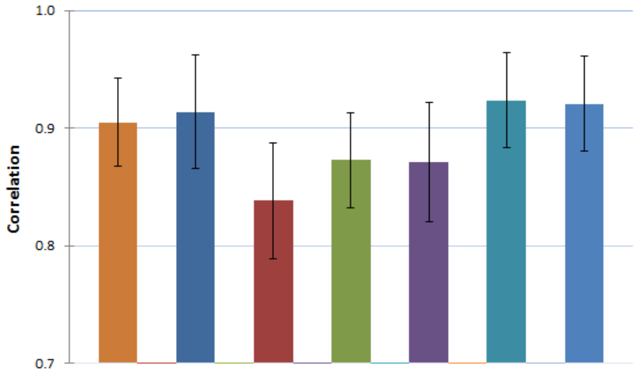
Fig. 7. Effect of the thresholds V_{bin} and V_{lin} on correlation between objective scores and subjective scores in the case of the binary and linear velocity response models.

TABLE II
SIZE AND AVERAGE COMPUTATION TIME FOR THE DIFFERENT REGRESSION METHODS

Regression method		Number of non-zero regression coefficients	Average computation time (s)
Single-stage regression		41 (binary model) 43 (linear model)	2354.11 (binary model) 2468.94 (linear model)
Two-stage regression	Stage 1	22 (binary model) 23 (linear model)	13.97 (binary model) 14.75 (linear model)
	Stage 2	8 (binary model) 7 (linear model)	4.77 (binary model) 3.37 (linear model)

First, to understand the effect of the choice of threshold value used in the velocity response function and to select an optimal value, the performances of the MoBin and MoLin methods are evaluated for different threshold values ranging from 1 to 10 pixels. For each value, the correlation between objective and subjective scores is used to measure the ability of the velocity response function to predict quality of experience based on motion alone. Results, shown in Fig. 7, indicate that a threshold of 3 pixel is optimal for both types of response functions and is therefore used in the rest of the paper.

Next, the performance of the different methods is evaluated by calculating the Pearson's Linear Correlation Coefficient (PLCC) between predicted scores and subjective scores over the entire set of test sequences. The performance of each method is shown in Fig. 8. Considering first the effect of the velocity response function, one can observe that the binary motion response model usually outperforms the linear model. The binary motion sensitive objective scores considered on their own appear to be better predictors than their linear counterpart as well as the non-motion sensitive objective scores as can be seen when comparing the performance of MoBin against MoLin and NoMo. This suggests that they are the most important type of objective scores. When combined with non-motion sensitive objective scores, the binary model remains a better predictor than the linear model, although the difference between the two becomes marginal. As for the effect of the regression method, it can be observed that the two-stage approach significantly improves performance compared to the single-stage approach for both types of velocity response functions. This can be attributed to improved convergence resulting from objective score selection. The detrimental effects of high dimensionality are evidenced by the poor performance of



	NoMo	MoBin	MoLin	ComBin	ComLin	SelBin	SelLin
Median	0.9222	0.9331	0.8407	0.8770	0.8757	**0.9402**	0.9395
Mean	0.9052	0.9137	0.8385	0.8728	0.8713	**0.9238**	0.9207
Stdv	**0.0592**	0.0764	0.0782	0.0643	0.0796	0.0643	0.0641
SE	0.0187	0.0241	0.0247	0.0203	0.0252	0.0203	0.0203
95% CI	0.0375	0.0483	0.0495	0.0406	0.0503	0.0406	0.0405

Fig. 8. Correlations between predicted scores and subjective scores for the different models considered with 95% confidence intervals.

the single-stage regression approaches (ComBin and ComLin) which perform more poorly than the non-motion sensitive model alone (NoMo).

Finally, Table II provides some information on the average size of the models built by the two regression approaches together with their computational time. The single-stage regression approach performs regression over 120 objective scores which results in models with large numbers of objective scores (over 40) and slow convergence (over 2000 s). In contrast, the two-stage approach enables selection of only a small number of coefficients (about 20) in the first stage which translates into significantly smaller final models (less than 10 coefficients) and reduces computation time by two orders of magnitude (less than 20 s for the combined two stages).

Overall, the models using either a binary or a linear velocity response function with a two-stage selective regression method are the top performers, resulting in the highest correlation scores while at the same time being significantly more compact and faster to compute. They are therefore the methods of choice that will be used in Section V-B to build the metrics.

B. Metric construction and evaluation

To build the metrics and evaluate their performances, the NAMA3DS1-CoSpaD1 dataset [47] is used for training and validation, while the ROMEO and Waterloo 3D-VQA datasets are used for testing. This ensures that there is no overlap between training and testing datasets and enables evaluation under diverse datasets covering a broad range of scenes and motion activity levels. The NAMA3DS1-CoSpaD1 dataset is split into a training set (9 sequences) and a validation set (1 sequence) to build different models. The model achieving the best performance on the validation set is used to build the metric. Metrics are then evaluated by calculating the Pearson's Linear Correlation Coefficient (PLCC), Spearman's Rank Correlation Coefficient (SRCC) and Kendall's Rank Correlation Coefficient (KRCC) between the predicted scores and the subjective scores over the testing datasets.

Two variants of motion sensitive metric are proposed depending on whether a binary or a linear velocity response function is used. These are evaluated against state-of-the-art quality metrics. More specifically the proposed metrics are compared against an image quality metric [28], a stereoscopic image quality metric [50], a video quality metric [51] and six stereoscopic video quality metrics [4], [8], [31], [33], [34]. The complete set of quality metrics evaluated and their key characteristics are listed as follows:

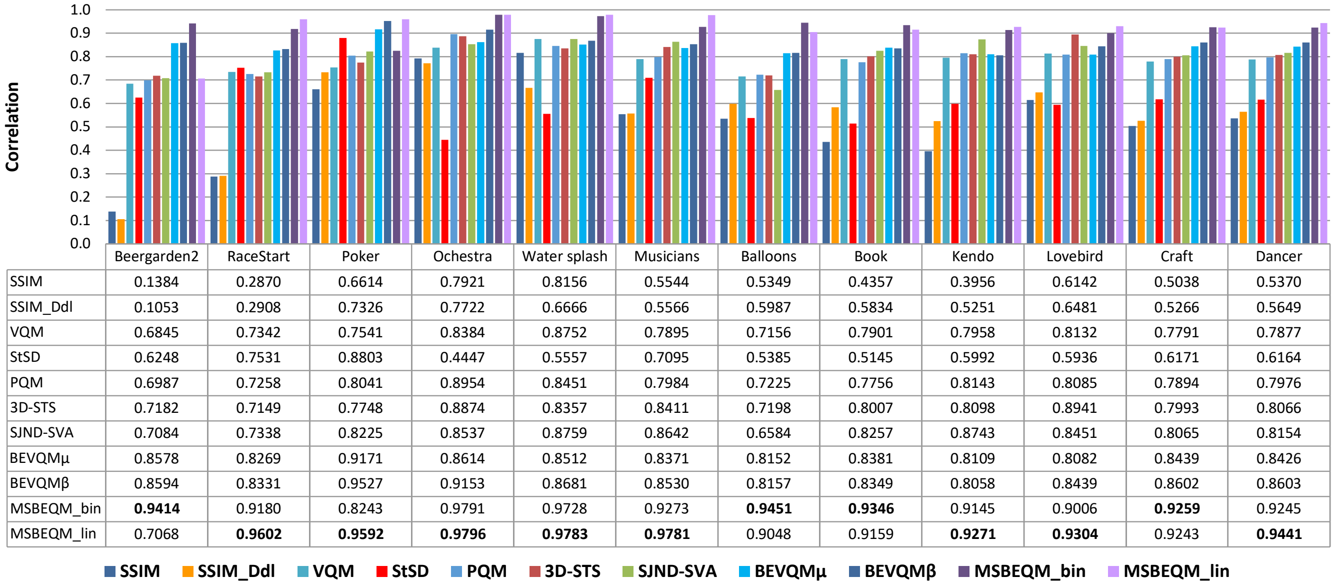
- **SSIM:** based on luminance, contrast and structural comparison [28],
- **SSIM_Ddl:** based on a global 2D image distortion measure and differences in disparity maps of stereo pairs [50],
- **VQM:** standardised method for objective evaluation of video quality [51],
- **StSD:** based on structural distortion, asymmetric blur and content complexity [4],
- **PQM:** based on distortions in luminance and contrast [31],
- **3D-STS:** based on spatio-temporal structure [33],
- **SJND-SVA:** based on visual attention and just-noticeable difference models [34],
- **BEVQMμ:** based on a non-motion sensitive HVS model with temporal pooling using averaging [8],
- **BEVQMβ:** same as above with temporal pooling using Minkowski summation [8],
- **MSBEQM$_{\text{bin}}$:** this is based on the proposed motion-sensitive HVS model with a binary velocity response function and two-stage regression,
- **MSBEQM$_{\text{lin}}$:** same as above with a linear velocity response function.

In the case of monoscopic quality metrics such as SSIM and VQM, stereoscopic quality scores are obtained by applying the quality metric separately to the left and right inputs and then averaging the left and right quality scores obtained. For the image quality metrics, mean temporal pooling is also performed to obtain a score for the entire video.

Fig. 9 shows the results obtained for the different quality metrics and for each sequence from the ROMEO and Waterloo 3D-VQA datasets based on PLCC. Further, Fig. 10 shows the average correlation on each dataset as well as the overall performance based on PLCC, SRCC and KRCC. In all cases, top performance has been highlighted in bold. Scatter plots showing the distribution of predicted scores against subjective scores are also provided in our supplementary report [49].

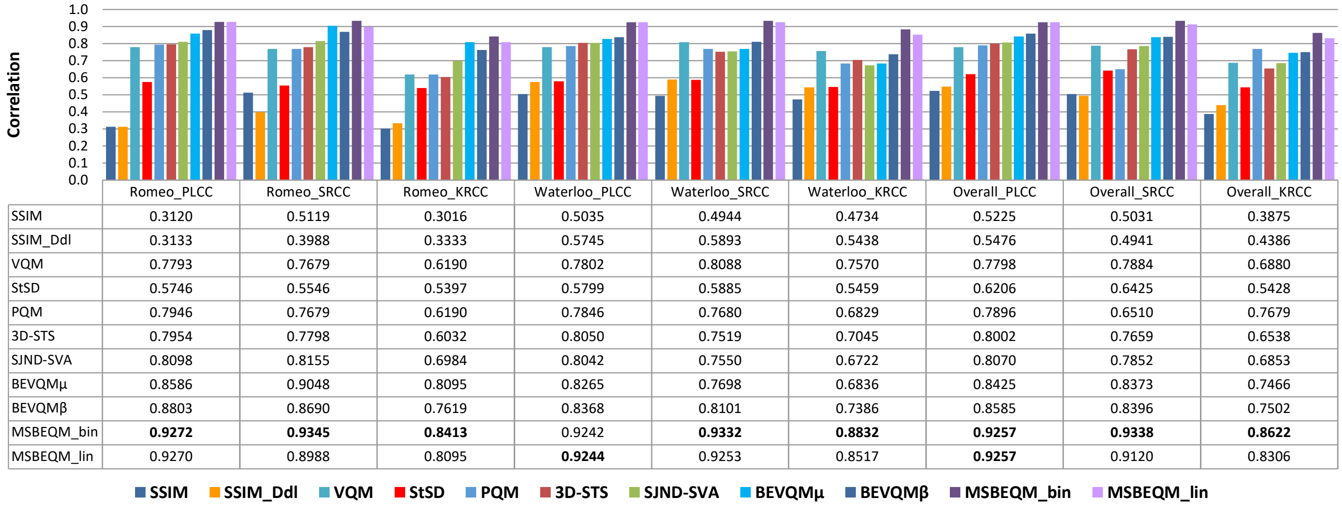
The image quality metrics, SSIM and SSIM_Ddl, are the two worst performers with average scores significantly lower than any video quality metric tested, whether stereoscopic or not. This confirms the importance of considering temporal information. The monoscopic video quality metric VQM performs less than the binocular video metrics considered, except for StSD which performs less well on these datasets. This demonstrates the benefit of accounting for binocular visual effects to achieve a high performance stereoscopic video quality assessment.

Considering now the performance of the different stereoscopic video quality metrics, it can be observed that the top four performers are all based on HVS models, being either



	Beergarden2	RaceStart	Poker	Ochestra	Water splash	Musicians	Balloons	Book	Kendo	Lovebird	Craft	Dancer
SSIM	0.1384	0.2870	0.6614	0.7921	0.8156	0.5544	0.5349	0.4357	0.3956	0.6142	0.5038	0.5370
SSIM_Ddl	0.1053	0.2908	0.7326	0.7722	0.6666	0.5566	0.5987	0.5834	0.5251	0.6481	0.5266	0.5649
VQM	0.6845	0.7342	0.7541	0.8384	0.8752	0.7895	0.7156	0.7901	0.7958	0.8132	0.7791	0.7877
StSD	0.6248	0.7531	0.8803	0.4447	0.5557	0.7095	0.5385	0.5145	0.5992	0.5936	0.6171	0.6164
PQM	0.6987	0.7258	0.8041	0.8954	0.8451	0.7984	0.7225	0.7756	0.8143	0.8085	0.7894	0.7976
3D-STS	0.7182	0.7149	0.7748	0.8874	0.8357	0.8411	0.7198	0.8007	0.8098	0.8941	0.7993	0.8066
SJND-SVA	0.7084	0.7338	0.8225	0.8537	0.8759	0.8642	0.6584	0.8257	0.8743	0.8451	0.8065	0.8154
BEVQMμ	0.8578	0.8269	0.9171	0.8614	0.8512	0.8371	0.8152	0.8381	0.8109	0.8082	0.8439	0.8426
BEVQMβ	0.8594	0.8331	0.9527	0.9153	0.8681	0.8530	0.8157	0.8349	0.8058	0.8439	0.8602	0.8603
MSBEQM_bin	**0.9414**	0.9180	0.8243	0.9791	0.9728	0.9273	**0.9451**	**0.9346**	0.9145	0.9006	**0.9259**	0.9245
MSBEQM_lin	0.7068	**0.9602**	**0.9592**	**0.9796**	**0.9783**	**0.9781**	0.9048	0.9159	**0.9271**	**0.9304**	0.9243	**0.9441**

Fig. 9. Performances of the proposed metrics against state-of-the-art metrics on each of the sequences from the ROMEO dataset (first 6 sequences) and the Waterloo 3D-VQA dataset (last 6 sequences) based on PLCC score.



	Romeo_PLCC	Romeo_SRCC	Romeo_KRCC	Waterloo_PLCC	Waterloo_SRCC	Waterloo_KRCC	Overall_PLCC	Overall_SRCC	Overall_KRCC
SSIM	0.3120	0.5119	0.3016	0.5035	0.4944	0.4734	0.5225	0.5031	0.3875
SSIM_Ddl	0.3133	0.3988	0.3333	0.5745	0.5893	0.5438	0.5476	0.4941	0.4386
VQM	0.7793	0.7679	0.6190	0.7802	0.8088	0.7570	0.7798	0.7884	0.6880
StSD	0.5746	0.5546	0.5397	0.5799	0.5885	0.5459	0.6206	0.6425	0.5428
PQM	0.7946	0.7679	0.6190	0.7846	0.7680	0.6829	0.7896	0.6510	0.7679
3D-STS	0.7954	0.7798	0.6032	0.8050	0.7519	0.7045	0.8002	0.7659	0.6538
SJND-SVA	0.8098	0.8155	0.6984	0.8042	0.7550	0.6722	0.8070	0.7852	0.6853
BEVQMμ	0.8586	0.9048	0.8095	0.8265	0.7698	0.6836	0.8425	0.8373	0.7466
BEVQMβ	0.8803	0.8690	0.7619	0.8368	0.8101	0.7386	0.8585	0.8396	0.7502
MSBEQM_bin	**0.9272**	**0.9345**	**0.8413**	0.9242	**0.9332**	**0.8832**	**0.9257**	**0.9338**	**0.8622**
MSBEQM_lin	0.9270	0.8988	0.8095	**0.9244**	0.9253	0.8517	**0.9257**	0.9120	0.8306

Fig. 10. Average performances of the proposed metrics against state-of-the-art metrics on the ROMEO dataset, the Waterloo 3D-VQA dataset and the combined ROMEO and Waterloo 3D-VQA datasets based on PLCC, SRCC and KRCC scores.

variants of the BEQM or the proposed BEVQM. Looking more closely at the performance of these four methods, it can be seen that the two variants of the proposed MSBEQM metric outperform their non-motion sensitive BEVQM counterparts by a significant margin. These results confirm the importance of modelling the motion sensitivity of the HVS when devising a stereoscopic video quality metric.

The two variants of the MSBEQM perform similarly with $\text{MSBEQM}_{\text{bin}}$ and $\text{MSBEQM}_{\text{lin}}$ achieving average correlations of 0.9257 (PLCC), 0.9338 and 0.9120 (SRCC), 0.8622 and 0.8306 (KRCC). This is in agreement with the results shown in Section V-A which suggested that the two models have similar performance. For the 'Beergarden2' sequence, a reduction in performance can be noted for $\text{MSBEQM}_{\text{lin}}$ compared to MSBEQMbin; this may be due to the high texture details present in the video which may lead to complex optical flows.

The complete list of non-zero coefficients for the $\text{MSBEQM}_{\text{bin}}$ and $\text{MSBEQM}_{\text{lin}}$ are given in Table III. These specify all the regression parameters defining the metric in (15). The listed coefficients specify the constant k and the non-zero entries of the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ identified by their indices. The coefficient indices in the range 1 to 60 refer to non-motion sensitive coefficients, while indices in the range 61 to 120 refer to motion sensitive coefficients. It can be observed that the $\text{MSBEQM}_{\text{bin}}$ and $\text{MSBEQM}_{\text{lin}}$ present some similarities in terms of coefficients that are selected with, in particular, the third motion sensitive objective score playing the most significant role in both models.

VI. Conclusions and Future Work

This paper introduced a motion sensitive HVS model based on physiological observations describing the response

TABLE III
COMPLETE LIST OF NON-ZERO COEFFICIENTS FOR THE $\mathrm{MSBEQM}_{\mathrm{BIN}}$ AND $\mathrm{MSBEQM}_{\mathrm{LIN}}$. ALL COEFFICIENTS ARE SORTED IN ORDER OF DECREASING IMPORTANCE BASED ON THEIR p-VALUE.

$\mathrm{MSBEQM}_{\mathrm{bin}}$				$\mathrm{MSBEQM}_{\mathrm{lin}}$			
Coeff.	Estimate	SE	p-value	Coeff.	Estimate	SE	p-value
k	4.348	0.12	0	k	3.927	0.10	0
a_{63}	-2.571	0.62	6.3E-5	a_{63}	-1.740	0.41	3.9E-5
$b_{3,63}$	-1.890	0.52	4.2E-4	a_3	-1.608	0.42	1.8E-4
a_{117}	1.169	0.32	4.4E-4	a_{114}	1.336	0.39	9.3E-4
a_{13}	-1.137	0.33	9.0E-4	$b_{3,114}$	-2.030	0.84	1.2E-2
a_{92}	1.764	0.56	2.1E-3				
a_{89}	-0.567	0.25	2.4E-2				
a_3	-0.621	0.46	1.8E-1				

of complex cells to motion. The approach is based on a generalised model of complex cells whose behaviour is defined by a velocity response function. Depending on the choice of velocity response function, the model is able to describe the behaviour of a wide range of complex cells including both non-motion sensitive and motion sensitive types. Although this paper only implemented the predominant type of motion sensitive complex cells, known to behave as high pass filters, the approach is generalisable to other types of complex cells.

This paper has demonstrated an application of the proposed model in stereoscopic content production. The model was used to define binocular energy terms capturing the non-motion sensitive and motion sensitive characteristics of each video frame. Temporal pooling and a two-stage regression approach were introduced to reduce dimensionality and improve the efficiency and accuracy of the estimation of a stereoscopic video quality metric. Two variants were proposed depending on whether a binary or a linear velocity response function is used to describe the behaviour of motion sensitive complex cells. Both metrics were evaluated on three stereoscopic video datasets containing a wide range of scenes and motion activity levels. The evaluation has showed that the two proposed metrics perform better than existing stereoscopic video quality metrics including other HVS-based metrics, and are able to achieve average correlations to subjective scores of 0.9257 (PLCC), 0.9338 and 0.9120 (SRCC), 0.8622 and 0.8306 (KRCC).

Further advances in understanding the physiology of the HVS are likely to open up new avenues to extend this research. For instance, a better understanding of the proportion of complex cells with motion sensitivity and more precise models of their velocity response functions would help increase the accuracy of the proposed approach. The present study assumed a common velocity threshold for all complex cells, however it may be beneficial to introduce cells with a variety of threshold values to better capture the effects of scene motion amplitude.

Furthermore, incorporating complex cells with different motion sensitivity responses such as low pass filter and band pass filter or even other types of non-linear velocity responses has the potential to increase the performance of the model and resulting metric. However, this is likely to also open up new computational challenges as the number of objective scores increases. Another interesting avenue for future research would be to extend the model by incorporating physiological findings modelling the response of other parts of the brain beyond the V1 area.

Another research direction would be to extend the approach by treating the stereoscopic video input as a 3D signal instead of two separate video streams, applying 3D image processing techniques such as the 3D transform to derive the objective scores. In this approach, motion sensitivity may be incorporated using scene flow instead of optical flow. Finally, it would be interesting to investigate the use of the proposed model in other application domains such as 3D video compression.

REFERENCES

[1] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *J. Vis. Commun. Image Represent*, vol. 22, no. 4, pp. 297–312, 2011.

[2] R. Bensalma and M.-C. Larabi, "A perceptual metric for stereoscopic image quality assessment based on the binocular energy," *Multidimens. Syst. Signal Process.*, vol. 24, no. 2, pp. 281–316, 2013.

[3] 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 Trans. Image Process.*, vol. 22, no. 5, pp. 1940–1953, 2013.

[4] 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 Trans. Image Process.*, vol. 22, no. 9, pp. 3392–3404, 2013.

[5] S. Ryu and K. Sohn, "No-reference quality assessment for stereoscopic images based on binocular quality perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 591–602, 2014.

[6] 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 *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2014, pp. 594–598.

[7] 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 *Proc. 3DTV Conf.*, 2015.

[8] C. Galkandage, J. Calic, S. Dogan, and J.-Y. Guillemaut, "Stereoscopic video quality assessment using binocular energy," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 102–112, 2017.

[9] Y. Lv, M. Yu, G. Jiang, F. Shao, Z. Peng, and F. Chen, "No-reference stereoscopic image quality assessment using binocular self-similarity and deep neural network," *Signal Process. Image Commun.*, vol. 47, pp. 346–357, 2016.

[10] F. Shao, W. Chen, G. Jiang, and Y.-S. Ho, "Modeling the perceptual quality of stereoscopic images in the primary visual cortex," *IEEE Access*, vol. 5, pp. 15 706–15 716, 2017.

[11] J. Ma, P. An, L. Shen, and K. Li, "Joint binocular energy-contrast perception for quality assessment of stereoscopic images," *Signal Process. Image Commun.*, vol. 65, pp. 33–45, 2018.

[12] K. Foster, J. Gaska, S. Marčelja, and D. Pollen, "Phase relationships between adjacent simple cells in the feline visual cortex," *J. Physiol.*, vol. 345, no. 1, p. 22P, 1983.

[13] D. A. Pollen and S. F. Ronner, "Phase relationships between adjacent simple cells in the visual cortex," *Science*, vol. 212, no. 4501, pp. 1409–1411, 1981.

[14] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106–154, 1962.

[15] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *JOSA A*, vol. 2, no. 2, pp. 284–299, 1985.

[16] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, and G. B. Akar, "Towards compound stereo-video quality metric: a specific encoder-based framework," in *Proc. IEEE Southwest Symp. Image Anal. Interpretation*, 2006, pp. 218–222.

[17] 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.

[18] 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 Res.*, vol. 32, no. 7, pp. 1193–1198, 1992.

[19] W. Waleszczyk, C. Wang, W. Burke, and B. Dreher, "Velocity response profiles of collicular neurons: parallel and convergent visual information channels," *Neuroscience*, vol. 93, no. 3, pp. 1063–1076, 1999.

[20] E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vision Res.*, vol. 38, no. 5, pp. 743–761, 1998.

[21] N. Qian and R. A. Andersen, "A physiological model for motion-stereo integration and a unified explanation of Pulfrich-like phenomena," *Vision Res.*, vol. 37, no. 12, pp. 1683–1698, 1997.

[22] Z.-L. Lu and G. Sperling, "The functional architecture of human visual motion perception," *Vision Res.*, vol. 35, no. 19, pp. 2697–2722, 1995.

[23] 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. Int. Workshop Video Process. Qual. Metrics Consum. Electronics*, 2010.

[24] P. Campisi, P. Le Callet, and E. Marini, "Stereoscopic images quality assessment," in *Proc. European Signal Processing Conference (EUSIPCO)*, 2007, pp. 2110–2114.

[25] R. G. Kaptein, A. Kuijsters, M. T. M. Lambooij, W. A. IJsselsteijn, and I. Heynderickx, "Performance evaluation of 3D-TV systems," in *Proc. SPIE Image Qual. Syst. Perform. V,*, vol. 6808, 2008.

[26] L. Goldmann, F. De Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," in *Proc. SPIE 3D Image Proc. Appl.*, vol. 7526, 2010.

[27] A. Tikanmaki, A. Gotchev, A. Smolic, and K. Miller, "Quality assessment of 3D video in rate allocation experiments," in *Proc. Int. Symp. Consum. Electronics*, 2008.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[29] L. Xing, J. You, T. Ebrahimi, and A. Perkis, "A perceptual quality metric for stereoscopic crosstalk perception," in *Proc. Int. Conf. Image Process.*, 2010, pp. 4033–4036.

[30] J. Yang, Y. Liu, Z. Gao, R. Chu, and Z. Song, "A perceptual stereoscopic image quality assessment model accounting for binocular combination behavior," *J. Vis. Commun. Image R.*, vol. 31, pp. 138–145, 2015.

[31] P. Joveluro, H. Malekmohamadi, W. C. Fernando, and A. Kondoz, "Perceptual video quality metric for 3D video quality assessment," in *Proc. 3DTV Conf.*, 2010, pp. 1–4.

[32] L. Jin, A. Boev, A. Gotchev, and K. Egiazarian, "3D-DCT based perceptual quality assessment of stereo video," in *Proc. Int. Conf. Image Process.*, 2011, pp. 2521–2524.

[33] J. Han, T. Jiang, and S. Ma, "Stereoscopic video quality assessment model based on spatial-temporal structural information," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process.*, 2012, pp. 1–6.

[34] F. Qi, D. Zhao, X. Fan, and T. Jiang, "Stereoscopic video quality assessment based on visual attention and just-noticeable difference models," *Signal Image Video Process.*, vol. 10, no. 4, pp. 737–744, 2016.

[35] K. Ha and M. Kim, "A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video," in *Proc. Int. Conf. Image Process.*, 2011, pp. 2525–2528.

[36] F. Lu, H. Wang, X. Ji, and G. Er, "Quality assessment of 3D asymmetric view coding using spatial frequency dominance model," in *Proc. 3DTV Conf.*, 2009, pp. 1–4.

[37] K. Seshadrinathan and A. C. Bovik, "Motion-tuned spatiotemporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, 2010.

[38] L. K. Choi and A. C. Bovik, "Video quality assessment accounting for temporal visual masking of local flicker," *Signal Process. Image Commun.*, vol. 67, pp. 182–198, 2018.

[39] I. W. Selesnick, R. G. Baraniuk, and N. G. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 123–151, 2005.

[40] J. Schanda, *Colorimetry: Understanding the CIE system*. John Wiley & Sons, 2007.

[41] G. Peyré and S. Mallat, "Orthogonal bandelet bases for geometric images approximation," *Commun. Pure Appl. Math.*, vol. 61, no. 9, pp. 1173–1212, 2008.

[42] I. M. Finn and D. Ferster, "Computational diversity in complex cells of cat primary visual cortex," *J. Neurosci.*, vol. 27, no. 36, pp. 9638–9648, 2007.

[43] 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." *J. Physiol.*, vol. 283, p. 101, 1978.

[44] L. Tao, M. Shelley, D. McLaughlin, and R. Shapley, "An egalitarian network model for the emergence of simple and complex cells in visual cortex," *Proc. Nat. Acad. Sciences*, vol. 101, no. 1, pp. 366–371, 2004.

[45] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. Scand. Conf. Image Anal.*, 2003, pp. 363–370.

[46] A. Ralston and H. S. Wilf, "Mathematical methods for digital computers," Tech. Rep., 1960.

[47] 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 *Proc. Int. Workshop Qual. Multimedia Experience*, 2012, pp. 109–114.

[48] J. Wang, S. Wang, and Z. Wang, "Asymmetrically compressed stereoscopic 3D videos: Quality assessment and rate-distortion performance evaluation," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1330–1343, 2017.

[49] C. Galkandage, J. Calic, S. Dogan, and J.-Y. Guillemaut, "Full-reference stereoscopic video quality assessment using a motion sensitive HVS model: Supplementary report," Tech. Rep., 2020.

[50] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *EURASIP J. Image Video Process.*, vol. 2008, p. 659024, 2009.

[51] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, 2004.

Chathura Galkandage is a Research Fellow in the Centre for Vision, Speech and Signal Processing at the University of Surrey and in the Department of Computer Science and Informatics at the London South Bank University. He received the BSc degree (first class) in Electronics and Telecommunications Engineering from the University of Moratuwa, Sri Lanka, in 2007 and the PhD degree from the University of Surrey, in 2017.

Janko Calic is a Visiting Lecturer at the Centre for Vision, Speech and Signal Processing, University of Surrey, and a Senior R&D Engineer at the BBC R&D. His main areas of expertise are QoE in video systems and user aspects of multimedia communications and HCI. 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 managed various EU-funded multinational collaborative research projects.

Jean-Yves Guillemaut is a Senior 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.