SUBJECTIVE EVALUATION OF 3D VIDEO ENHANCEMENT ALGORITHM

Federica Battisti, Marco Carli, and Alessandro Neri

Department of Engineering
Universita' degli Studi Roma TRE
Roma, Italy

ABSTRACT

In this contribution the subjective evaluation of a 3D enhancement algorithm is presented. In the proposed scheme, perceptually significant features are enhanced or attenuated according to their saliency and to the masking effects induced by textured background. In particular, for each frame we consider the high frequency components, i.e., the edges, as relevant features in the edge complex wavelet domain computed by the first order dyadic Gauss-Laguerre Circular Harmonic Wavelet decomposition. The saliency is assessed by evaluating both disparity map and motion vectors extracted from the 3D videos. The effectiveness of the proposed approach has been verified by means of subjective tests.

Index Terms— Video enhancement, stereo, subjective quality, Laguerre Gauss

1. INTRODUCTION

The release of Avatar, which grossed $2.7 billion worldwide, has been the seed for industry to move towards the three-dimensional technology. On one hand, content producers were particularly attracted by 3D world as a media for sharing information and improving the quality of experience, on the other hand, content providers and manufacturers, i.e., movie theaters, 3D televisions, 3D game consoles, smartphones, and other devices, devised the opportunity for new revenues from three-dimensional entertainment.

Despite of the big investment the 3D format has yet to take off. Consumers were barely persuaded to trade in their televisions, buy new cameras or new game consoles. And, even if they did, the percentage of people really using the 3D technology is very low. Discomfort, fatigue, and visual stress caused by the low quality of the rendering system and by the abuse of effects looking at amplifying immersivity in 3D spaces maybe cited as strong motivation for the low popularity of 3D, as demonstrated in [1–4]. Despite the efforts spent in understanding the Human Visual System (HVS) and the characteristics of the 3D perception, and in modeling the related Quality of Experience [5, 6], the improvement of user acceptance and satisfaction based on the characteristics of 3D human visual perception is at an early stage of development [7]. The simplest way of performing 3D image restoration and enhancement, consists on separately operating on each view of the stereo pair, using algorithms developed for the monocular case [8]. Those approaches result in feeling of artificial clarity. To cope with this problem, methods exploiting the characteristics of the HVS in the enhancement process have been presented. In [9] the authors locally increase the 3D image contrast based on depth information. Similarly, in [10] sharpness and contrast of 3D videos are adjusted based on the disparity information of objects extracted from 3D videos.

In this contribution we propose a novel 3D video enhancement technique that aims at providing a visual experience similar to the one experimented when looking at natural scenes. It is based on a space-variant multiresolution image enhancement driven by the information collected from the spatial and temporal organization of the objects in the observed scene. The goal of our approach is in the application of the "Organization of Space" criteria in which figures appearing in the foreground should possess well-defined contours, while the background should be more undetermined, being at a far distance from the viewer [11, 12].

2. PROPOSED APPROACH

In order to exploit the Organization of Space of a natural scene, we combine the information provided by the depth information on each image element, derived for instance from a disparity map, with the optic laws providing quantitative evaluation of the bandwidth reduction of an optical signal versus the traveled distance. The spatio-temporal Contrast Sensitivity Function (CSF) shows that moving stimuli (which are not being tracked and therefore would have non-zero retinal velocity) are perceived more sharply than non-moving targets with a bandwidth characteristic. In our approach we adopt a simplification of the spatio-temporal CSF accounting for the higher difficulty for the HVS in perceiving details of a fast moving object with respect to a still one.

The processing is performed in the wavelet domain. Here, we have adopted the Laguerre Gauss Circular Harmonic Wavelets (LG) [13, 14]. The Gauss-Laguerre functions belong to the class of the CHFs, widely employed in rotation invariant pattern recognition [15], harmonic tomographic

decomposition [16], and rotation-invariant pattern signatures [17]. The rationale for our choice stems from the fact that, for any given resolution, the corresponding scalogram is essentially a complex map of elementary stimuli constituting the image like edges, lines, forks, and corners, whose magnitude represents the strength of a feature at a given point, while its phase is proportional to its orientation. Thus, restoration and enhancement respectively reduce to a shrinking, and to an amplification of the magnitude of each wavelet coefficient, based on a rule that depends on the depth and on the motion vector associated to the corresponding image site.

Given a stereo pair $f^m(\mathbf{x}), m = L, R$, its enhancement can be performed by jointly selectively amplifying the higher scales (LG) expansion coefficients. As common practice in multiresolution analysis, we decompose the generic frame into a coarse approximation $f^m_{s_0}(\mathbf{x})$ obtained by applying to the original image a low pass zero order LG filter, and the details $\Delta f^m_{s_k}(\mathbf{x})$ at finer scales obtained by filtering $\Delta f^m(\mathbf{x}) = f^m(\mathbf{x}) - f^m_{s_0}(\mathbf{x})$ with the scaled versions of higher order LG filters. Formally, we compute the enhanced version $\tilde{f}^m(\mathbf{x})$ as follows:

$$\tilde{f}^{(m)}(\mathbf{x}) = \sum_{k=1}^{L} \Delta\tilde{f}^{(m)}_{s_k}(\mathbf{x}) * g_{s_k}(\mathbf{x}) + f^{(m)}_{s_0}(\mathbf{x}), \tag{1}$$

where:

$$\Delta\tilde{f}^{(m)}_{s_k}(\mathbf{x}) = \eta_{s_k}\left[\Delta f^{(L)}_{s_k}(\mathbf{x}), \Delta f^{(R)}_{s_k}(\mathbf{x}), z(\mathbf{x}), v(\mathbf{x})\right], \tag{2}$$

η_{s_k} is a pointwise (memoryless) function that depends on both detail strengths $\Delta f^{(m)}_{s_k}$, the detail depth $z(\mathbf{x})$, and the magnitude of the motion field $v(\mathbf{x})$, and $g_{s_k}(\mathbf{x})$ are the LG reconstruction filters.

Let us assume that the stereo pair has been rectified in order to comply with an epipolar geometry. In this case, assuming coplanar pinhole cameras with parallel optical axes, the disparity maps $d^{(L)}(\mathbf{x})$ represent the horizontal displacement between the corresponding points of the left with respect to the right frame, and is inversely proportional to the object depth $z(\mathbf{x})$. Thus, denoting with $O^{(L)}$ the set of points of the left frame without corresponding point in the right frame (e.g. points belonging to occluded areas) and with $\bar{O}^{(L)}$ its complement set, we have:

$$f^L(x_1, x_2) = f^R\left[x_1 + d^{(L)}(x_1, x_2), x_2\right], \forall (x_1, x_2) \in \bar{O}^{(L)}. \tag{3}$$

The same relation holds for the right frame. Then

$$\begin{aligned} d^{(L)}(x_1, x_2) = -d^{(R)}[x_1 + d^{(L)}(x_1, x_2), x_2] \\ \forall (x_1, x_2) \in \overline{O^{(L)} \cup O^{(R)}}. \end{aligned} \tag{4}$$

The disparity map and the magnitude of the motion field are then combined in order to control the amount of enhancement performed over different regions of a given frame pair. Let $|d|_{\min}$ and $|d|_{\max}$ respectively be the minimum and the maximum disparity magnitude for a given imaging geometry, and v_{max} the maximum magnitude of motion vector compatible with the expected object dynamics. Let us define the normalized left and right disparity maps $d^{(m)}_0(\mathbf{x})$ as follows:

$$d^{(m)}_0(\mathbf{x}) = \begin{cases} \frac{|d^{(m)}(\mathbf{x})| - |d|_{\min}}{|d|_{\max} - |d|_{\min}} & |d|_{\min} \le |d^{(m)}(\mathbf{x})| \le |d|_{\max} \\ 1 & |d^{(m)}(\mathbf{x})| > |d|_{\max} \end{cases}. \tag{5}$$

Similarly, the normalized magnitude of the motion field $v^{(m)}_0(\mathbf{x})$ are defined as:

$$v^{(m)}_0(\mathbf{x}) = \begin{cases} \frac{v^{(m)}(\mathbf{x})}{v_{\max}} & v^{(m)}(\mathbf{x}) \le v_{\max} \\ 1 & v^{(m)}(\mathbf{x}) > v_{\max} \end{cases} \tag{6}$$

Then, for each pair the normalized enhancement factor maps $e^{(m)}_0(\mathbf{x})$ are built by linearly combining the normalized disparity maps and the normalized motion field magnitudes, i.e.,

$$e^{(m)}_0(\mathbf{x}) = \alpha \cdot [1 - d^{(m)}_0(\mathbf{x})] + (1 - \alpha) \cdot [1 - v^{(m)}_0(\mathbf{x})]. \tag{7}$$

A greater enhancement factor is associated to those objects which are the slowest, and closest to the observer. The parameter controls the relative relevance of the objects' spatial and temporal behavior. We remark that the enhancement factor is identical for corresponding points of the left and right image. In fact the enhancement factors associated to the right and the left frame, satisfy the relationship:

$$\begin{aligned} e^{(L)}_0(x_1, x_2) = e^{(R)}_0\left[x_1 + d^{(L)}(x_1, x_2), x_2\right], \\ \forall (x_1, x_2) \in \overline{O^{(L)} \cup O^{(R)}}. \end{aligned} \tag{8}$$

Therefore, Equation 1 becomes:

$$\tilde{f}^{(m)}(\mathbf{x}) = \sum_{k=1}^{L} \gamma_{s_k}\left[e^{(m)}_0(\mathbf{x})\right] \Delta f^{(m)}_{s_k}(\mathbf{x}) * g_{s_k}(\mathbf{x}) + f^{(m)}_{s_0}(\mathbf{x}). \tag{9}$$

In our experiments the design of the relationship between γ_{s_k} and $e^{(m)}_0(\mathbf{x})$ has been inspired to the relationships between the signal bandwidth B and the link length L observed in optical communications for which we have:

$$B = \frac{B_0}{L^{\delta}} \tag{10}$$

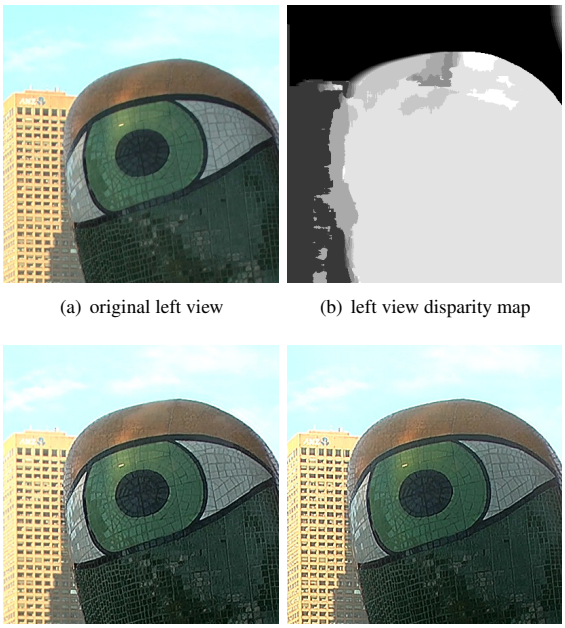
where $0.5 \le \delta \le 1$ depending on the light dispersion nature. Thus the gain coefficients γ are selected in such a way that the spatial bandwidth of the reconstruction filter is proportional to $\left[e^{(m)}_0(\mathbf{x})\right]^{-\delta}$. At this aim, we observe that if a constant set

of gains $\{\gamma_{s_k},\ k = 1, ..., L\}$ is employed, the overall transfer function is

$$G_{tot}(\omega) = \gamma_{s_1} \frac{\sum_{k=1}^{L} \frac{\gamma_{s_k}}{\gamma_{s_1}} \left| \mathcal{F}\left\{ \frac{1}{s_k} \mathcal{L}_0^{(1)} \left(\frac{r(\mathbf{x})}{s_k}, \theta(\mathbf{x}) \right) \right\} \right|^2}{\sum_{k=1}^{L} \left| \mathcal{F}\left\{ \frac{1}{s_k} \mathcal{L}_0^{(1)} \left(\frac{r(\mathbf{x})}{s_k}, \theta(\mathbf{x}) \right) \right\} \right|^2}. \quad (11)$$

The corresponding bandwidth versus $\{\gamma_{s_k},\ k = 1, ..., L\}$ can then be easily computed. We remark that the bandwidth is controlled only by the relative magnitude of the gains (i.e. magnitude normalized w.r.t. one of them), their absolute magnitude affects the degree of edge sharpness produced by the procedure. An example of the 3D image enhancement based on the above criterion is illustrated in Figure 1 where a detail of the left image of the first frame pair of the shot 3D_25 of the RMIT uncompressed stereoscopic 3D HD video library [18] and its processed versions are reported.

Visual comparison of Figure 1.a and Figure 1.c reveals that the finer and richer details of the foreground object have been enhanced. At the same time buildings in the foreground exhibit a small blurring with respect to the original, with the exception of a small region on the foreground object contour. This last effect is caused by the disparity estimation algorithm employed. In fact, as illustrated in Figure 1.b, the area with large disparity is thicker than the foreground object, due to the large size (i.e. 21 pixels) of the window used by the block matching algorithm employed in the disparity map estimation. For sake of comparison, in Figure 1.c and in Figure 1.d the version obtained by applying the enhancement in the complex edge domain without spatial bandwidth adaptive control is also reported. As expected, in this case, background and foreground are enhanced in an undifferentiated manner. As a consequence the enhanced image appears not natural. These considerations are consistent with the results presented in [6] achieved results are consistent with.



(a) original left view (b) left view disparity map

(c) enhanced view with spatial bandwidth adaptive control (d) enhanced view without spatial bandwidth control

Fig. 1. Detail of example of frame enhancement based on the use of the disparity map.

3. SUBJECTIVE EVALUATION

In this Section, the performances of the proposed method are presented and discussed.

3.1. Methodology

Equipment: The subjective test was conducted at the Istituto Superiore delle Comunicazioni e delle Tecnologie dell'Informazione 3D quality test laboratory. The laboratory setup had controlled lighting system to produce reliable and repeatable results. The evaluation was performed using a 46" Hyundai S465D polarized stereoscopic monitor with a native resolution of 1920x1080 pixels. *Observers*: twenty-four naive viewers (6 female and 18 male), with a marginal experience of 3D image and video viewing, evaluated the quality of each test sequence. The age distribution ranged from 23 to 52. The viewers were seated at a distance of about four times the height of the active part of the display ($\sim$ 170 cm). All subjects underwent a screening to examine their visual acuity, color vision, and stereo vision. *Stimuli*: to generate the test sequences, a set of six stereo video sequences, side by side, progressive, of size 1920x1080 pixels and varying frame rate in the range 25-30fps, representing sequences of duration 10 seconds each were chosen [19]. The selected sequences present different video content and scene motion rate: cartoon (Dracula and Knights Quest), computer generated scene (Dzignlight and Peschke), and natural scenes (Treffen and Skydiving) characterized by slow and fast content change. The sequence Knights has been used for training purposes.

3.2. Procedure

We adopt the standard subjective evaluation methodology for 3D video quality assessment [20]. In the subjective experiments, the Single Stimulus (SS) evaluation system, was adopted. During the test, subjects were presented with one stereo video sequence independently rated. The sequences were presented in random order.

After the presentation of each sequences, a five seconds time interval for voting followed. The evaluation was based on a 5-level judgment test. The scores are corresponding to:

excellent (5), good (4), fair (3), poor (2), and bad (1). Each test session was composed by five stages: instruction, training, practice trials, experimental trials, and final interview. In the first stage, the subject was verbally given instructions about the experiment procedure. He/she was made familiar with the stereo viewing setup and with the experiment graphical interface. In the training stage, the test interface showing the original video, and samples of test video were used to familiarize the subject with the experiment procedure. Following, to stabilize subjects' responses, practice trials were performed with one video with 5 different enhancing rates. These sequences are not then used in the real test and the subjects are informed about this. Finally, in the interview stage, information about the overall annoyance or interest in the topic was collected.

3.3. Data analysis

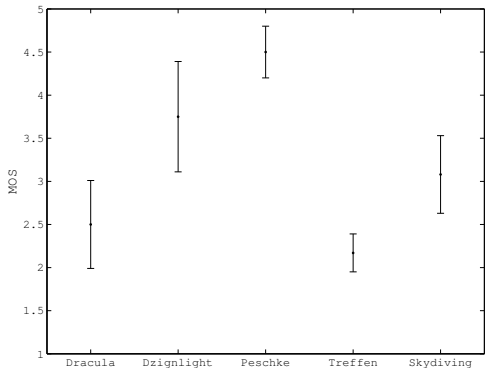
The screening of the subjects was performed according to the guidelines described in ITU-R BT.500-11. No subjects were excluded. For each video the Mean Opinion Score (MOS) has been computed $MOS_k = \overline{u_k} = \frac{1}{N}\sum_{i=1}^{N} u_k^{(i)}$ where $u_k^{(i)}$ is the i^{th} subject score for the k^{th} sequence and N represents the total number of observers. For evaluating the role of sampling error in the performed estimation, the 95% - confidence interval δ_k is computed, as described in [21].

In Figure 2 the MOS values and the relative confidence interval level δ_k are reported. As can be noticed, the average confidence interval is around 0.44 thus showing a good estimation performance. In Figure 3 the increase in MOS score resulting from the enhancement proposed scheme is shown.

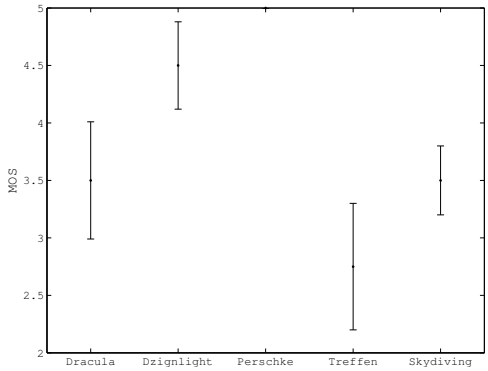
The subjective scores agree with the subjective comments collected during the experiment. The general remark was of more natural feeling when looking at some videos (corresponding to the enhanced versions). In particular all subjects reported a positive comment on the enhanced version of the video *Peschke* describing the *naturalness* feeling. In Figure 4 sample left frames from the original and the enhanced version of this video, together with the relative disparity map, are reported.

4. CONCLUDING REMARKS

A 3D Video Enhancement driven by the perception of spatial organization by the HVS has been proposed. In the presented approach, for exploiting the Organization of Space of a natural scene, the information provided by the depth information on each image element are fused with the optic laws providing quantitative evaluation of the bandwidth reduction of an optical signal versus the traveled distance. Furthermore, the higher difficulty for the HVS in perceiving details of a fast moving object with respect to a still one, is considered.



(a) Original video.



(b) Enhanced video.

Fig. 2. MOS for the original and the enhanced videos.

REFERENCES

[1] K. Yamagishi, L. Karam, J. Okamoto, and T. Hayashi, "Subjective characteristics for stereoscopic high definition video," in *Quality of Multimedia Experience (QoMEX), Int. Work. on*, 2011, pp. 37–42.

[2] L. Stelmach, J.T. Wa, D. Meegan, and A. Vincent, "Stereo image quality: effects of mixed spatio-temporal resolution," *Circuits and Systems for Video Technology, IEEE Trans. on*, vol. 10, no. 2, pp. 188–193, 2000.

[3] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "New requirements of subjective video quality assessment methodologies for 3DTV," *Video Processing and Quality Metrics, (VPQM)*, 2010.

[4] M. Lambooij, M. Fortuin, W. IJsselsteijn, and I. Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: A review," *Jour. of Imaging Science and Technology*, vol. 53, no. 3, pp. 1–14, 2009.

[5] K. Wang, M. Barkowsky, K. Brunnstrom, M. Sjostrom, R. Cousseau, and P. Le Callet, "Perceived 3D TV Transmission Quality Assessment: Multi-Laboratory Results Using Absolute Category Rating on Quality of Experi-

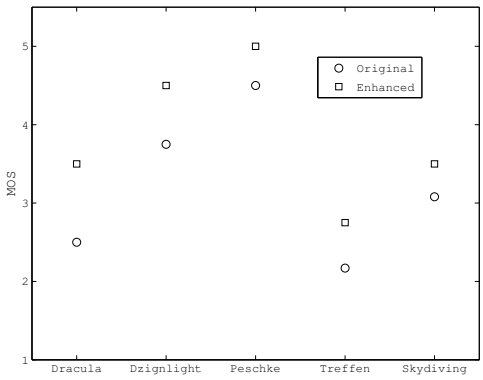
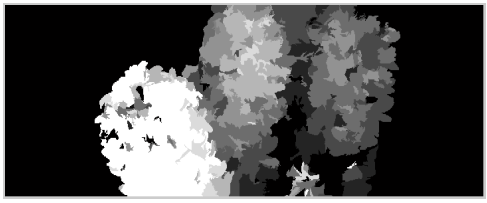


Fig. 3. MOS for the original and the enhanced videos.

(a) Original video *Peschke*.

(b) Enhanced version.



(c) Disparity map.

Fig. 4. Original, enhanced, and disparity map of a sample frame extracted from the *Peschke* sequence.

ence Scale," *IEEE Transactions on Broadcasting*, vol. 58, no. 4, pp. 544–557, 2012.

[6] J. Wang, M. Barkowsky, V. Ricordel, and P. Le Callet, "Quantifying how the combination of blur and disparity affects the perceived depth," in *Proc. SPIE Electronic Imaging*, 2011, vol. 7865.

[7] X. Cao, A.C. Bovik, Y. Wang, and Q. Dai, "Converting 2d video to 3d: An efficient path to a 3d experience," *MultiMedia, IEEE*, vol. 18, no. 4, pp. 12–17, 2011.

[8] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image restoration by sparse 3D transform-domain collaborative filtering," *Procs. SPIE Image Processing: Algorithms and Systems VI*, vol. 6812, 2008.

[9] W. Hachicha, A. Beghdadi, and F. A. Cheikh, "Combining depth information and local edge detection for stereo image enhancement," 20^{th} *European Signal Processing Conf.*, pp. 250 – 254, 2012.

[10] B. Govem, M. Sayinta, E. Somcag, and F. Donmez, "Depth based 3D sharpness and contrast enhancement application on stereo images," *Procs. of* 21^{st} *Signal Processing and Communications Applications Conf.*, pp. 1–4, 2013.

[11] A. Neri, P. Campisi, E. Maiorana, and F. Battisti, "3D Video Enhancement based on Human Visual System characteristics," *Procs. of International Workshop on Video Processing and Quality Metrics (VPQM)*, 2010.

[12] A. Neri, P. Campisi, and F. Battisti, "Fuzzy Edge Enhancement in the Complex Wavelet Domain," *Procs. of International Workshop on Video Processing and Quality Metrics (VPQM)*, 2009.

[13] L. Costantini, L. Capodiferro, M. Carli, and A. Neri, "Texture segmentation based on laguerre gauss functions and k-means algorithm driven by kullbackleibler divergence," *Journal of Electronic Imaging*, vol. 22, no. 4, pp. 043015–043015, 2013.

[14] A. Neri and G. Jacovitti, "Maximum likelihood localization of 2-d patterns in the gauss-laguerre transform domain: theoretic framework and preliminary results," *IEEE Trans. on Image Processing*, vol. 13, no. 1, pp. 72–86, 2004.

[15] H. H. Arsenault and Y. Sheng, "Properties of the circular harmonic expansion for rotation invariant pattern recognition," *Applied Optics*, vol. 25, no. 18, pp. 3225–3229, 1986.

[16] S. R. Deans, *The Radon Transform and Some of Its Applications*, Wiley, New York, 1983.

[17] E. P. Simoncelli, "A rotation invariant pattern signature," *Procs. of Third IEEE International Conference on Image Processing*, vol. 3, pp. 185–188, 1996.

[18] "RMIT 3D," http://www.rmit3dv.com/.

[19] "3D TV," http://3dtv.at/.

[20] "Subjective methods for the assessment of stereoscopic 3dtv systems," *International Telecommunication Union ITU-R BT.2021*, 2012.

[21] "Methodology for the subjective assessment of the quality of television pictures," *International Telecommunication Union ITU BT.510-11*, 2002.