3VQM: A 3D Video Quality Measure

Mashhour Solh, *Member, IEEE,* Ghassan AlRegib, *Senior Member, IEEE,* Judit Martinez Bauza, *Member, IEEE*.

Abstract

Three dimensional television (3DTV) and free viewpoint television (FTV) are believed to be the future of television broadcasting that would bring a more life-like and visually immersive home entertainment experience. These technologies will provide the users with the freedom to navigate through the scene and choosing a desired viewpoint. A desired viewpoint video can be synthesized at the receiver side using depth image-based rendering (DIBR). In this paper, we present a new method for objectively evaluating the quality of stereoscopic 3D videos generated by DIBR. The quality of the synthesized 3D videos is affected by errors or noise in the depth maps. To evaluate the distortions in DIBR-based 3D videos a reference depth is needed for evaluation. In this paper, we show how to derive an ideal depth estimate at each pixel value that would constitute a distortion-free rendered video. The ideal depth estimate will then be used to derive three distortion measures to objectify the visual discomfort in the stereoscopic videos. The three measures are defined as temporal outliers (`TO`), temporal inconsistencies (`TI`), and spatial outliers (`SO`). The combination of the three measures will constitute a visual quality measure for 3D DIBR-based videos, `3VQM`. `3VQM` can be both derived in a full-reference (`FR-3VQM`) and no-reference (`NR-3VQM`) scenarios. Finally, the proposed measure will be verified and validated against a fully conducted subjective evaluation and compared to 2D-based quality measures used in literature to evaluate 3D video quality. The results show that our proposed measure is significantly accurate, coherent and consistent with the subjective scores and outperforms the few 2D-based metrics proposed in literature.

Index Terms

Video Quality, View Synthesis, Stereoscopic Quality Assessment, Depth-Based Image Rendering(DIBR) ,3D Video, 3DTV

M. Solh and G. AlRegib are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: {msolh, alregib}@gatech.edu. J. Martinez Bauza is a staff engineer with Multimedia R&D, Qualcomm, San Diego, CA, USA e-mail: juditm@qualcomm.com.

I. INTRODUCTION

In the last few years 3D videos have witnessed a significant surge in popularity. Many movies nowadays are being recorded and/or presented in a stereoscopic 3D format. This trend is accompanied with noticeable advances in stereoscopic and autostreoscopic display technologies for 3DTV and mobile devices. For these reasons, it is widely believed that 3D video will overtake current high definition 2D video as the future of multimedia broadcasting. In the future, 3D video experience will be more life-like and visually immersive. This is also known as free-viewpoint TV (FTV) [1]. FTV users will have the freedom to navigate through the scene to choose a different viewpoint. In current stereoscopic 3D video systems, each individual viewpoint requires two videos corresponding to the left and right camera views. Hence, to capture and broadcast an arbitrary viewpoints for 3D display in FTV an unrealistic number of cameras will be required. It will also require extremely complex and efficient coding, and expensive computing capabilities. In addition, advances in 3D display technologies require a flexibility in the number of views (autostereoscopic displays) and the ability to resize each view to match the display resolution. Consequently, the impractical volume of resources needed to capture sufficient number of views opens the door for more innovative ways to generate a large number of views via interpolation between views using view synthesis. Several techniques have been proposed in the literature for view synthesis [2]. Among these techniques, depth image-based rendering (*DIBR*) has drawn much attention for generating views for FTV and 3D videos in both simply and efficiently [3]. In DIBR, two or more views for 3D display can be generated from a single 2D image and a corresponding depth map using 3D wrapping [3]. Advantages in DIBR include but are not limited to bandwidth-efficiency, interactivity by synthesizing virtual views from various view points, easy 2D to 3D switching, and computational and cost efficiency hence less cameras are needed. Moreover, DIBR eliminates photometric asymmetries in between the two views hence both of them are generated from the same original image. These advantages have lead MPEG to issue a standard for coding DIBR format or MPEG-C part 3 [4]. The synthesized views in DIBR are generated using 3D wrapping which first projects the pixels in the reference image back to the world coordinates using explicit geometric information form the depth map and camera parameters, the resulting pixels in the world coordinates are then projected back to the estimated virtual image coordinate [5].

Holes appear when occluded areas in the reference image becomes visible in the virtual image. Holes could also result from wrong or noisy depth values. As a result some image processing or hole-filling is required to fill in these hole areas. In section III we will review view synthesis using DIBR. A DIBR-

based 3DTV system processing chain (depicted in Fig. 1) is composed of six main components [3] [6]:(i) 3D video capturing and content generation;(ii) 3D content video coding; (iii) transmission;(iv) decoding the received sequences;(v) generating virtual views; (vi) displaying the stereoscopic images on the screen.

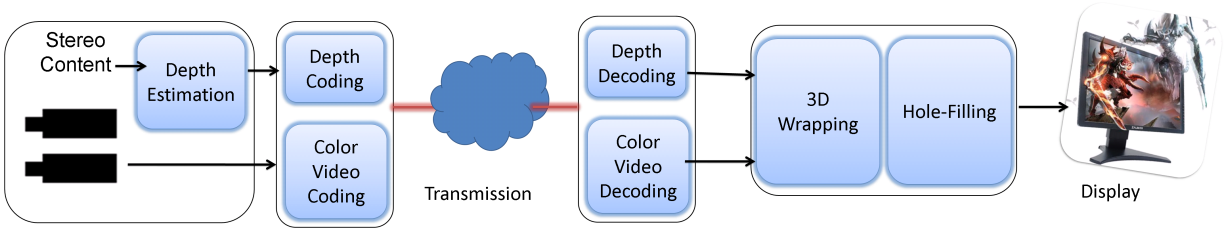


Fig. 1. Block diagram of the DIBR-based 3DTV processing chain.

The perceived quality of the DIBR-based stereoscopic 3D videos is sensitive to every block in the processing chain. The end user's perceived quality will be affected by the following:

- accuracy of the estimated depth maps,
- quality of the 3D wrapping process in DIBR,
- quality of the hole-filling algorithm applied to cover the disoccluded areas in the generated frames,
- compression artifacts for the 2D video and depth map,
- transmission errors and streaming losses, and
- scaling and formatting algorithms in the 3D displays.

In this paper, we introduce an new objective visual quality measure for DIBR-based 3D videos: `3VQM: 3D Video Quality Measure`. `3VQM` estimates elements of the visual discomfort in DIBR synthesized stereoscopic videos based on the *ideal depth* estimate. The *ideal depth* is a new concept that we define as the per pixel depth that would generate a DIBR-based distortion-free 3D video. In this paper, we will explain the *ideal depth* and show how to derive it in a full-reference and no-reference scenarios. While 2D video quality is solely based on monocular cues in one view (texture, color, blur, blocking artifacts ...), 3D video quality on the other hand is a combination of binocular and monocular cues. The depth illusion in stereoscopic video is constructed by presenting the eye with two views with slight horizontal disparity or binocular cues. Depth is also perceived through a number of monocular cues such as lighting, shading, motion parallax, texture gradient, blur, relative sizes, and occlusion. The importance of each one these cues for depth perception may vary depending on the scene. Visual discomfort occurs whenever the depth through monocular cues mismatches or conflicts the depth through binocular cues.

Errors and artifacts introduced in the the processing pipeline of DIBR 3D videos (Fig. 1) could result

in conflicts in the depth cues such as unmatched color objects, mismatches between the blur in different depth planes and disparity, unmatched luminance and frame cancelation (near-edge cut-off for objects with front depth). In addition, errors in depth map that could result from inaccurate estimation, numerical rounding, or compression artifacts. These errors may lead to distortions in the relative pixel location and in the magnitude of the pixels. The visual effect of these distortions on the synthesized view is *spatially noticeable* around texture areas in the form of significant intensity changes and *temporally noticeable* around flat regions in the form of flickering [7]. Visual discomfort may also arise from other factors such as *excessive disparities, fast changing disparities*, and *geometric distortions* in stereoscopic 3D videos [8].

In this paper the three distortions measures that will be introduced evaluate the temporal and spatial variation of the depth errors for the depth values that would lead to inconsistencies between the left and right view, fast changing disparities, and geometric distortions. These measures are the spatial outliers (`SO`), temporal outliers (`TO`), and temporal inconsistencies (`TI`). We package all these distortion measures into a new full-reference and no-reference visual quality measure for DIBR-based 3D videos: `FR-3VQM` and `NR-3VQM` respectively. The proposed measures are evaluated against subjective scores and compared against contemporary quality measurement techniques.

The rest of the paper is organized as follows. The related Work will be reviewed in Section II. View synthesis using DIBR will be reviewed in Section III. Then the concept of *ideal depth* estimation will be presented in Section IV. In Section V we will present the three distortion measures to measure visual discomfort in stereoscopic videos. The combination of three measures constituting our vision-based quality measure for 3D DIBR-based videos will be introduce in Section VI. Finally, the experimental results and conclusion will be presented in Sections VII and VIII, respectively.

II. Related Work

The most common technique to evaluate stereoscopic 3D videos is subjective evaluation. Research efforts have been dedicated to evaluate parameters that would influence the subjective quality such as display size, camera configuration, viewing distance and positioning [9]–[12]. These research efforts are accompanied with standardization efforts for subjective evaluation by the International Telecommunication Union (ITU) and the Video Quality Experts Group (VQEG). ITU has issued some recommendations for subjective methods for assessment of stereoscopic 3DTV systems in Recommendations $BT.500-11$ (in 2002) and $P.910$ (in 2008), but these recommendations are outdated and limited to picture and depth quality. VQEG has been focusing on the subjective evaluation for cross talk in 3DTV systems and is

currently working on a test plan for 3D quality assessment [13].

Objective quality assessment for stereoscopic 3D videos is a recent research topic and the existing measures are relatively few. The majority of the current objective stereoscopic video quality metrics methodologically can be categorized as an extension to existing 2D metrics [14]–[17]. These techniques follow a simplistic approach by calculating the 2D quality measure of the left and the right image separately and then finding the combination of the values that would best predict the 3D video quality. These methods assume that the perceived depth distortions are less significant than the perceived color distortions. In [18] the authors proposed using *PSNR* and *SSIM* for 3DTV video as a quality measure for video plus depth content by measuring the quality of the virtual views that are rendered from the distorted color and depth sequences. The undistorted reference sequence is obtained by rendering virtual views from the original color and depth maps. Approaches based on 2D metrics have poor correlation with perceived 3D video quality and have been proved to be non-robust [19]. Other works in the literature include a no-reference measure based on evaluating the blockiness and disparity temporally, and then finding the best combination of parameters using particle swam optimization [20]. No depth information is considered in the aforementioned measure and it suffers from the same robustness and poor correlation problem of 2D video quality-based techniques. Ozbek et al. [21] assumed that PSNR of the second view is less important for 3D visual experience and the new measure was composed of weighted combination of two PSNR values and a jerkiness measure for temporal artifacts. An objective metric for free-viewpoint video production was proposed in [22]. The metric can be used as full-reference measure of fidelity of aligning structural details in presence of approximate scene geometry of the 3D shapes.

While all the aforementioned techniques ignored the depth information, authors in [23] used a combination of a depth-map error-based comparison function and 2D quality measure for colored images to predict the 3D image quality. Similar approaches were adapted by the authors in [24] and [25]. The addition of the depth information to the combination did not result in a significant improvement to the prediction of the 3D video quality. This can be attributed to the fact that visual discomfort was not considered in analyzing depth information. Among the most recent objective quality measures, a measure based on disparity, disparity-gradient maps, and spatial image activity was proposed in [26]. Similarly, the authors in [27] proposed a model for deriving overall quality of experience from image and depth quality. Finally, a measure for visual fatigue was also proposed in [28] based on the distributions of horizontal, vertical and angular pixel disparities. These measures considered the quality of synthesized 3D videos using depth based rendering but not the multitude of variables that would result in visual discomfort. Among these variables are excessive disparities, fast changing disparities, geometric distortions, temporal

flickering and spatial noise in the form of depth cues inconsistency. In contrast, 3VQM is a quality measure for synthesized stereoscopic videos generated by DIBR that takes these variables into consideration and as a result *3VQM* correlates well with subjective scores. The main component of 3VQM is the *ideal depth* that is presented in section IV. First, we will review the view synthesis process using DIBR in the following section.

III. View Synthesis

View synthesis using DIBR is composed of two main components: 3D wrapping and hole-filling (Fig. 1.

A. 3D Wrapping

In DIBR, virtual views are generated by first projecting the pixels in the reference image to the world coordinates using depth map and camera information. The resulting pixels in the world coordinates are then sampled in the 2D plane from a different viewpoint to obtain a DIBR estimated image. This process is known as 3D wrapping [5].

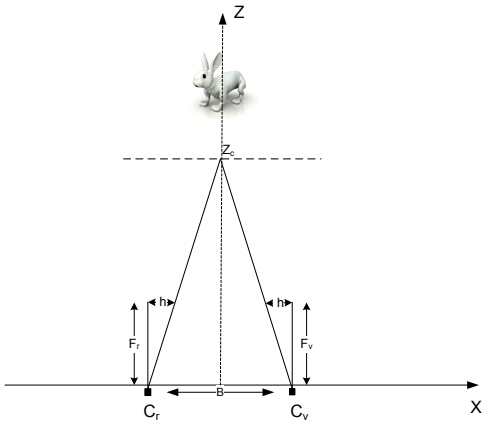


Fig. 2. Depth image-based rendering (DIBR).

The setup of view synthesis using DIBR is illustrated in Fig. 2. Let C_r be the reference camera and C_v be the virtual camera. The view corresponding to the virtual camera can be estimated from the view of the reference camera as follows:

$$\overline{X}_v = \overline{X}_r + s\frac{F_v B}{\overline{Z}} + h, \tag{1}$$

where $\overline{X}_v$ is the horizontal coordinates vector of the virtual camera, and $\overline{X}_r$ is the horizontal coordinate vector of the reference camera. F_r and F_v are the focal lengths of the reference and the virtual cameras,

respectively. [1] B is the baseline distance that separates the two cameras and Z_c is the convergence distance of the two cameras. $s = -1$ when the estimated view is to the left and $s = +1$ when the estimated view is to the right, $\overline{Z}$ is a vector of the depth values at pixel location (x_r, y_r), and h is the horizontal shift in the camera axis. h can be estimated as follows:

$$h = -s\frac{FB}{Z_c}. \tag{2}$$

B. Hole-Filling

Disocclusion occurs when occluded areas in the reference image become visible in the virtual image. Dissocclusion is depicted in the the black regions in the examples shown in Fig. 3. The resulting pixels in the disoccluded areas are also known as holes. Holes may also be caused by errors in depth map such as bad pixels from stereo matching. Disocclusion removal or hole-filling is a challenging problem that has received considerable research in recent years. Proposed hole-filling solutions include interpolation after depth map smoothing, extrapolation, background mirroring [3], asymmetric smoothing [29], distance dependent smoothing [30], edge-based smoothing [31], layered depth images (LDI) [32], and hierarchical hole filling (HHF) [33]. For more information on hole-filling we refer the reader to [33].

(a) (b)

Fig. 3. Disocclusion: (a) `Art` before hole-filling (b) `Aloe` before hole-filling.

IV. Ideal Depth Estimation

Video quality assessment can be classified into a full-reference, reduced reference and no-reference quality measures. In a 2D video full-reference scenario both the original video sequence from the sender

[1] F_r and F_v will be assumed to be equal to F for the rest of this paper.

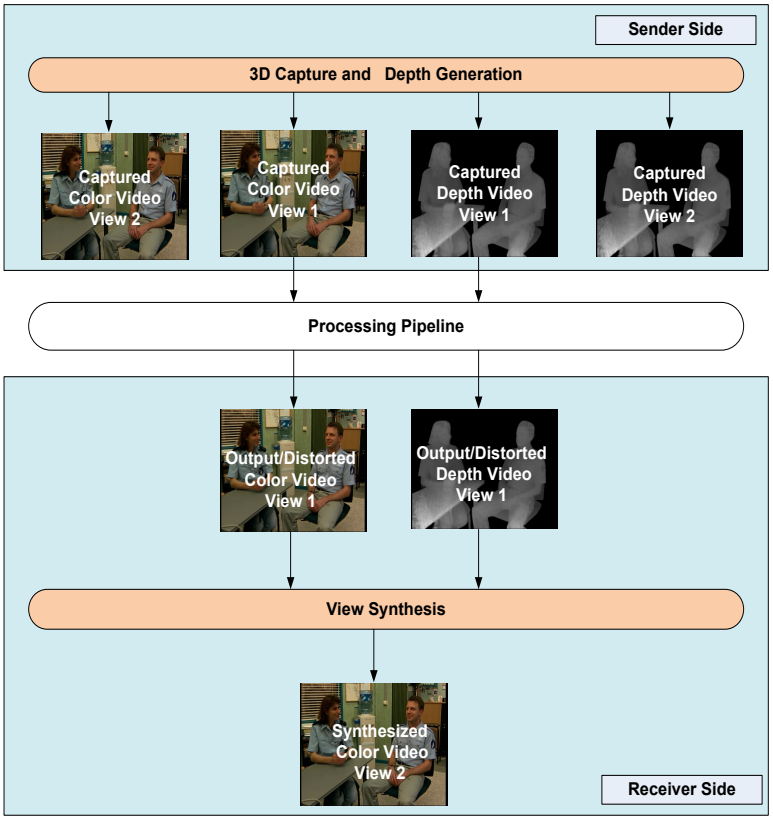


Fig. 4. In an DIBR based setting the depth is usually captured by an active sensor such as a time of flight(TOF) sensor by a passive sensor such as stereo matching in a multi-camera setting.

side and the corresponding processed video sequence at the receiver side are available for evaluation. In such cases, there is an implicit assumption that the original video sequence at the sender side is distortion free. Similarly, in a full-reference quality assessment for DIBR-based stereoscopic 3D video (as shown in Fig. 4) both captured color videos for stereoscopic views (captured color video `view 1` and `view 2` in Fig. 4) and one depth map (captured depth video for `view 1` in Fig. 4) from the sender side are available for evaluation. Given these videos, the quality of DIBR-based stereoscopic 3D could be measured by evaluating one or more of the following:

1) the distortions in the synthesized color video at the receiver side (synthesized color video `view 2` in Fig. 4) as compared to the corresponding view at the sender side (captured color video of `view 2` in Fig. 4),
2) the distortions in the received or processed color video at the receiver side (output/distorted color video `view 1` in Fig. 4) as compared to the corresponding captured view at the sender side

(captured color video of `view 1` in Fig. 4),

3) and the distortions in the received or processed depth at the receiver side (output/distorted depth video `view 1` in Fig. 4) as compared to the generated depth video at the sender side (captured depth video of `view 1` in Fig. 4).

In using any of the above options, there is an implicit assumption that the captured color and/or depth videos at the sender side are distortion free. This assumption might be valid for the color videos given that these videos are captured by high quality cameras. Nevertheless, this assumption is not accurate because it has to be defined within what constitute as an acceptable quality of stereoscopic experience. Notably, defining what qualifies subjectively as a good stereoscopic experience is still an ongoing research work.

As for depth, it is neither valid nor accurate to assume that the generated depth is distortion free. The current depth video capturing or sensing technologies are noisy, inaccurate and unreliable [34]. Depth can be either captured using a passive sensor that extracts depth by disparity estimations using stereo matching techniques, or an active sensor such as time-of-flight (TOF) camera. Passive sensors are particularly inaccurate around non-textured and featureless regions because it lacks visual information which makes it hard to establish correspondence across the views of multiple cameras. Active sensors on the other hand has a very low resolution and tends to be very noisy around textured regions [34]. As a result, the captured depth cannot be a valid reference for quality evaluation because the noises introduced by the capturing device adds up to the quality degradation by other sources of noise such as wrong estimations, numerical rounding and compression artifacts introduced by the processing pipeline.

In this paper, we define quality of experience by the amount of visual discomfort that the stereoscopic video might cause to the observer. Visual discomfort in synthesized stereoscopic videos using DIBR is mainly caused by noise in depth map. Depth map noise usually leads to inaccurate relocation of pixels during the wrapping process. Henceforth the resulting synthesized videos may suffer from excessive disparities, fast changing disparities, geometric distortions, temporal flickering and/or spatial noise in the form of depth cues inconsistency.

In order to quantify the visual discomfort that is caused by noise or errors in the depth map, we need a noise-free depth map as a reference. An ideal depth map reference for quality assessment is the per pixel depth that would generate a distortion-free virtual view using the same reference image and same DIBR parameters. We will refer to this depth as the *ideal depth*. A conceptual illustration for *ideal depth* is shown in Fig. 5. The *ideal depth* would make an excellent reference for our quality evaluation because it meets the following properties:

- The *ideal depth* is free of the noises introduced by the capturing devices and the processing pipeline.

- The *ideal depth* generates a distortion free synthesized color video using DIBR.
- The *ideal depth* is estimated from the captured color video. Therefore, the *ideal depth* is a valid reference to evaluate non depth related distortions such as distortions caused by the hole-filling algorithm and/or the colored video compression.

In the following subsections we will show how to estimate the *ideal depth* in full-reference and in no-reference case.

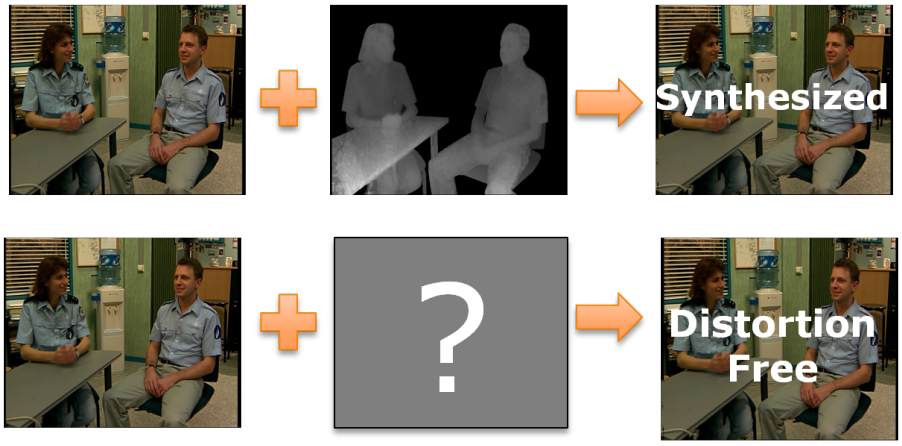


Fig. 5. *Ideal depth* is the depth map that would generate the distortion-free image given the same reference image and DIBR parameters (B, s, F_v, h).

A. Ideal Depth Estimation in the full-reference Case

The *ideal depth* estimation in the full-reference case is a function of the captured color video for the view to be interpolated (captured color video for `view 2` in Fig. 4) from the sender side, the received depth map (output/distorted depth video for `view 1` in Fig. 4) and the synthesized color video (synthesized color video for `view 2` in Fig. 4). The ideal depth estimate can be derived as follows:

1) Using the 3D wrapping equation in (1), we first express the horizontal coordinate $\bar{X}_v$ vector of the synthesized virtual view as a function of the horizontal coordinate vector of the reference view $\bar{X}_r$:

$$\bar{X}_v = \bar{X}_r + s\frac{F_v B}{\bar{Z}} + h \tag{3}$$

2) Similarly, the horizontal coordinate vector of the captured view $\bar{X}_o$ can be expressed as a function of the horizontal coordinate vector $\bar{X}_r$ of the the reference view:

$$\bar{X}_o = \bar{X}_r + s\frac{F_v B}{\bar{Z}_{IDEAL}} + h \tag{4}$$

where $\bar{Z}_{IDEAL}$ is the *ideal depth* map vector to be estimated. The distortion free view is assumed to be the captured color video (captured color video for `view 2` in Fig. 4).

3) By subtracting (4) from (3) and then preforming direct substitution, the *ideal depth* vector $\bar{Z}_{IDEAL}$ can be expressed as:

$$\bar{Z}_{IDEAL} = \frac{sF_vB}{(\bar{X}_o - \bar{X}_v) + s\frac{F_vB}{Z}} \tag{5}$$

4) Calculating $(\overline{X}_o - \overline{X}_v)$ in equation (5) is non-trivial. However, calculating the intensity variation $(\overline{I}_o - \overline{I}_v)$ is simpler and produces more accurate results than the horizontal shift $(\overline{X}_o - \overline{X}_v)$. Hence to estimate $\overline{Z}_{IDEAL}$ we need to derive a relationship between the intensity variation and the horizontal shift. In [7] the relation between the sum of squared differences (SSD) of the original video frame and its horizontal translations has been shown to be linear. Based on this observation we were able to prove that the horizontal shift values for each pixel location can be estimated in terms of the intensity variations as follows $\Delta\bar{I} \approx \alpha\Delta\bar{X}$ for a small horizontal shift $\Delta\bar{X}$, where α is a constant. The proof is given in section IV-C. The *ideal depth* now can be estimated from the rendered virtual view intensity vector $\bar{I}_v$, the distortion-free view intensity vector $\bar{I}_o$, the received depth map $\bar{Z}$ vector, focal length F_v, and the baseline B as follows:

$$\bar{Z}_{IDEAL} \approx \frac{sF_vB}{\alpha(\bar{I}_o - \bar{I}_v) + s\frac{F_vB}{Z}} \tag{6}$$

Now that we have derived the ideal depth estimate for the full-reference case, we will show how to derive it in the no-reference case.

B. Ideal Depth Estimation in the no-reference Case

The *ideal depth* estimation for the no-reference case is different from the full-reference case. In a no-reference case no information from the sender is available for evaluation. Therefore, the derivation for the no-reference *ideal depth* estimate proceeds as follows:

1) Equation (6) is the full-reference *ideal depth* estimate. For the no-reference case $\bar{I}_o$ is not available for evaluation. Therefore, we cannot explicitly derive the *ideal depth* map. Hence, we need to estimate the intensity variation vector $(\bar{I}_o - \bar{I}_v)$ from the intensity vector of the rendered virtual image $\bar{I}_v$ and the intensity vector $\bar{I}_r$ of the received reference image (synthesized color video for `view 2` and output/distorted depth video for `view 1` in Fig. 4) . If we assume this function to be $f(\bar{I}_v, \bar{I}_r)$, then the *ideal depth* can be expressed as a function of $f(\bar{I}_v, \bar{I}_r)$ as follows:

$$\bar{Z}_{IDEAL} \approx \frac{sF_vB}{\alpha(f(\bar{I}_v, \bar{I}_r)) + s\frac{F_vB}{\bar{Z}}}. \quad (7)$$

2) The intensity vector $\bar{I}_r$ of the received reference image (output/processed color video for `view 1`) is the closest in computational features to $\bar{I}_o$ among the available videos at the receiver side. However, before calculating the intensity variation for equation (7) we need to correct for the horizontal disparity between $\bar{I}_r$ and $\bar{I}_o$. Hence, the function $f(\bar{I}_v, \bar{I}_r)$ is calculated as the difference in intensity between each block in the reference view $\bar{I}_r$ and the corresponding block in the rendered virtual view $\bar{I}_v$ after applying a horizontal shift to the blocks of the reference view. $\bar{Z}_{IDEAL}$ can then be calculated in an algorithmic manner as shown in Algorithm 1.

Algorithm 1 Ideal depth approximation.

d is variable initialized as the block size

for $i = 1$ to $imagewidth$ step d **do**

  for $j = 1$ to $imageheight$ step d **do**

    D = Z[i to $i + d$, j to $j + d$]

    m = **mean**(D)

    $Iref$ = I_r[i to $i + d$, $j + m$ to $j + d + m$]

    $Iver$ = I_v[i to $i + d$, j to $j + d$]

    f [i to $i + d$, j to $j + d$] = $Iref$ - $Iver$

    Z_{IDEAL} [i to $i + d$, j to $j + d$] = $(sF_vB)/(\alpha(f) + s\frac{F_vB}{D})$

  end for

end for

The choice of the block size d does affect the noise level in the estimated *ideal depth*. This effect will be discussed in the results section.

C. Relationship Between Small Intensity Change and Small Horizontal Shift

In [7] the relation between the sum of squared differences (SSD) of the original video frame and its horizontal translations has been shown to be linear. In this section we will prove that the relationship between $\Delta\bar{X}$, the small horizontal shift values for each pixel location, and $\Delta\bar{I}$, the intensity or luminance variations, can be expressed as $\Delta\bar{I} \approx \alpha\Delta\bar{X}$, where α is a constant. The intensity or luminance of an image can be expressed as function of the horizontal and vertical coordinates $(\bar{X}, \bar{Y})$ as follows:

$$\bar{I} = f(\bar{X}, \bar{Y}). \tag{8}$$

But since we are only looking for variations along the horizontal coordinates $\bar{I}$ can be expressed in terms of $\bar{X}$ only where $\bar{Y}$ is assumed to be fixed, as follows:

$$\bar{I} = f(\bar{X}). \tag{9}$$

If we apply a Taylor series expansion of (9) in the neighborhood of 0, then (9) can be written as follows:

$$I = f(\bar{X}) = f(0) + \frac{f'(0)}{1!}\bar{X} + \frac{f''(0)}{2!}\bar{X}^2 + \frac{f^{(3)}(0)}{3!}\bar{X}^3 + ... \tag{10}$$

For a small change $\Delta\bar{X}$, the intensity at $\bar{X} + \Delta\bar{X}$ can be expressed as follows:

$$f(\bar{X} + \Delta\bar{X}) = f(0) + \frac{f'(0)}{1!}(\bar{X} + \Delta\bar{X}) + \frac{f''(0)}{2!}(\bar{X} + \Delta\bar{X})^2 + \frac{f^{(3)}(0)}{3!}(\bar{X} + \Delta\bar{X})^3 + ... \tag{11}$$

Subtracting (11) from (10) yields the following:

$$f(\bar{X} + \Delta\bar{X}) - f(\bar{X}) = \frac{f'(0)}{1!}(\Delta\bar{X}) + \frac{f''(0)}{2!}(2\bar{X}\Delta\bar{X} + (\Delta\bar{X})^2) + ... \tag{12}$$

The equation in (12) can be then reduced to the following:

$$\Delta\bar{I} = \frac{f'(0)}{1!}(\Delta\bar{X}) + \frac{f''(0)}{2!}(2\bar{X}\Delta\bar{X} + (\Delta\bar{X})^2) + ... \tag{13}$$

In order to derive the relationship between $\Delta\bar{I}$ and $\Delta\bar{X}$ we need to inspect the two coefficients that multiplies $\Delta\bar{X}$ and $(2\bar{X}\Delta\bar{X} + (\Delta\bar{X})^2)$ in (13). We ran a set of simulations on a database of stereoscopic images and videos. For each image or video frame we calculated ($\frac{f'(0)}{1!}$ and $\frac{f''(0)}{2!}$). In Fig. 6 and Fig. 7 we have chosen the plots for four images and video frames that consist a variation in depth and texture distributions. These sequences are the *Pantomime* and *Cafe* video sequences [35], the *Ballet* sequence [36], and the *Art* sequence [37]. The *Pantomime* video sequence has a medium complex depth and largely smooth texture structure. The *Cafe* video sequence has larger depth distribution and medium complex texture structure.The *Ballet* video sequence has a complex depth and smooth texture structure. The *Art* image sequence has a complex depth and complex texture structure.

For the plots in Fig. 6 and Fig. 7 we calculated $\frac{f'(0)}{1!}$ and $\frac{f''(0)}{2!}$ for each image and video frame. The results show that $\frac{f'(0)}{1!}$ is much larger than $\frac{f''(0)}{2!}$. We may also infer from the plots that for a small $\Delta\bar{X}$

the term $(\frac{f''(0)}{2!}(2\bar{X}\Delta\bar{X} + (\Delta\bar{X})^2)$ is very small compared to $\frac{f'(0)}{1!}(\Delta\bar{X})$ and hence the former can be assumed to be zero. In the figures of Fig. 6 and Fig. 7 we plotted the two terms for least ($\Delta\bar{X} = 1$) and most ($\Delta\bar{X} = 16$) against the horizonal coordinate X, where X is confined to a vector of size 16. The gradient values in here are for the middle rows of the images.

As a result, equation (13) can be reduced to the following:

$$\Delta\bar{I} \approx \frac{f'(0)}{1!}(\Delta\bar{X}), \tag{14}$$

which is also the linear approximation of $\Delta\bar{I}$. It is then valid to assume $\Delta\bar{I} \approx \alpha\Delta\bar{X}$, where α is a constant. The latter statement does actually mean that for small shifts along the horizontal axis the change of intensity tends to be proportional to the shift. This statement is true for most natural images, with small exception that could manifest around sharp edges.

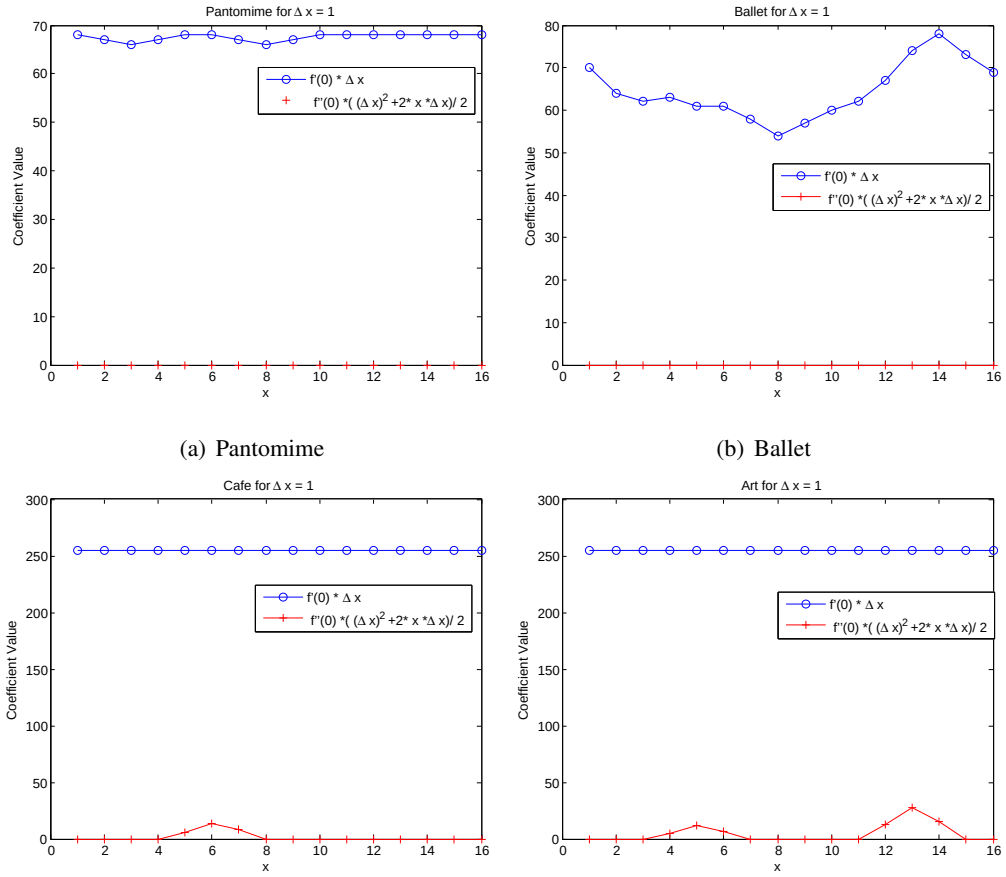
V. Distortion Metrics

Up to this point, we have derived an estimation of the *ideal depth*. In what follows, we will use the *ideal depth* to derive the distortion metrics that would account for visual discomfort in the synthesized video. We start by defining the term $\Delta\mathbf{Z}$, as the difference between the *ideal depth* and the received depth which can be expressed as follows:

$$\Delta\mathbf{Z} = |\mathbf{Z}_{IDEAL} - \mathbf{Z}|. \tag{15}$$

When the value of $\Delta\mathbf{Z}$ is zero at a certain pixel location that means that the corresponding pixel location is distortion free. However,a non-zero value of $\Delta\mathbf{Z}$ does not necessarily mean a visible distortion at that pixel location. For instance, a consistent (uniform) error over a specific depth plane will cause the whole plane to be shifted in one direction and the perceptual effect of such error will be a slight increase or decrease in the perceived depth. This slight increase or decrease does not constitute a perceptible visual distortion. These inaccuracies are spatially uniform and originate from estimations in the wrapping equation and inherited approximation in the camera modeling parameters. Otherwise, a non-zero value of $\Delta\mathbf{Z}$ does constitute a visual distortion in the synthesized video. Such visual distortions are the sources of the visual discomfort experienced by the end user.

To measure the visual discomfort we define three distortion metrics: the spatial outliers (`SO`), temporal outliers (`TO`), and temporal inconsistencies (`TI`).



(a) Pantomime (b) Ballet

(c) Cafe (d) Art

Fig. 6. Plot the two terms of equation (13) for least ($\Delta\bar{X} = 1$) horizontal variations against the horizonal coordinate X, where X is confined to a vector of size 16.

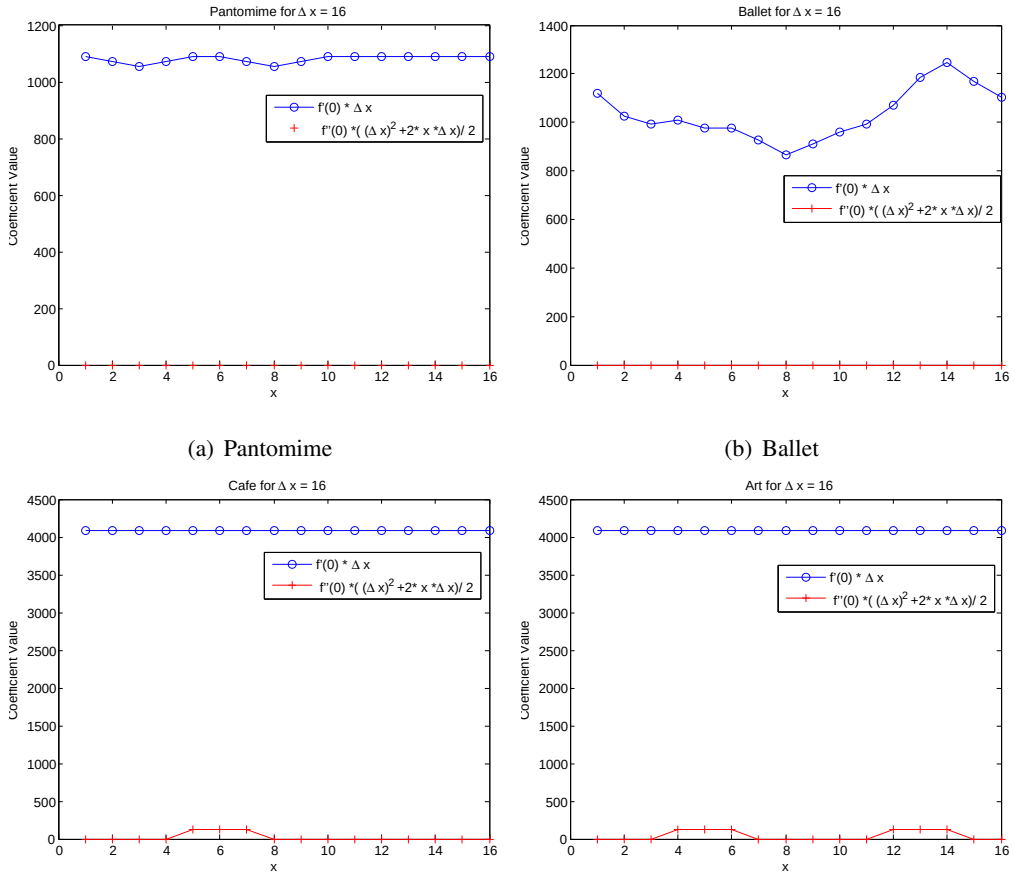
A. *Spatial Outliers (SO)*

A non-zero set of values of $\Delta\mathbf{Z}$ with non-uniform distribution over a depth plane would result in relocation of color pixel/blocks during the wrapping process to an alien position. The visual effect of these errors on the synthesized view is *spatially noticeable* around texture areas and would result in visual discomfort in the form of *inconsistent depth cues* (unmatched object colors) and *geometric distortions*.

These spatial inconsistencies can be quantified through the spatial outliers (**SO**), calculated as the standard deviation of $\Delta\mathbf{Z}$:

$$\mathbf{SO} = STD(\Delta\mathbf{Z}) \tag{16}$$

The standard deviation in this case separates the spatially visible distortions due to non-zero $\Delta\mathbf{Z}$ from



(a) Pantomime (b) Ballet

(c) Cafe (d) Art

Fig. 7. Plot the two terms of equation (13) for most ($\Delta\bar{X} = 16$) horizontal variations against the horizonal coordinate X, where X is confined to a vector of size 16.

the perceptually non-significant $\Delta\mathbf{Z}$'s. In Fig. 8 a frame from a DIBR generated video is shown. The original stereo video was captured by Point Grey's Bumblebee2 camera and then the depth map sequence was generated using stereo matching. The depth was then used to obtain a DIBR-based estimate of the right view video. Looking at the chosen frame we can see that there are distortions around the hand, the paper, the head, and the wall in background. These distortions are caused by both the errors in the depth maps as well as by the hole filling algorithm. The **SO** map of the frame in Fig. 8 is shown in Fig. 9(a). The spatial distortions were all captured by **SO** plus the edges where a plan shift occurs. The latter is not a source of visual distortion however it can be filtered using the temporal outliers described next.

B. Temporal Outliers (TO)

The temporal variation of $\Delta\mathbf{Z}$ is also another indicator of visual distortion resulting in visual discomfort. A temporally inconsistent $\Delta\mathbf{Z}$ indicates random pixel relocation during the wrapping process or inconsistency in the hole filling algorithm. This is spatially noticeable around textured areas in the form of *significant intensity changes* and around flat regions in the form of *flickering*. Therefore, we define the temporal outliers (**TO**) metric as the standard deviation of the change in $\Delta\mathbf{Z}$ for two consecutive frames:

$$\mathbf{TO} = STD(\Delta\mathbf{Z}_{t+1} - \Delta\mathbf{Z}_t) \tag{17}$$

The error introduced due to depth map noise is temporally inconsistent while an non-zero $\Delta\mathbf{Z}$ around at an edge of plane change will be temporally consistent because the same wrapping parameters were used to generate both frames. By taking the standard deviation the temporal outliers filters out the edginess in **SO** and will only keep the visible temporal distortions from depth map errors and hole-filling. This can be also observed by looking into the **TO** map of the Fig. 8, shown in Fig. 9(b) where the edginess is no longer part of the captured distortion.

C. Temporal Inconsistencies (TI)

Excessive disparities and *fast changing disparities* are another source of visual discomfort and are mainly caused by errors in stereo matching, hole-filling algorithms and depth compression. These distortions are also observed in the form of flickering, which is usually observed around smoothly textured areas and noise around highly structured regions. We will refer to this measure as the temporal inconsistencies metric (**TI**) and it can be derived as:

$$\mathbf{TI} = STD(\mathbf{Z}_{t+1} - \mathbf{Z}_t) \tag{18}$$

The **TI** map of the frame in Fig. 8 is shown in Fig. 9(c). We can notice that **TI** captures all the flickering on the wall in the background. This flickering is caused by inconsistencies in the hole filling algorithm. **TI** also captures the fast changing noises that were not captured by the spatial outliers earlier.

VI. 3VQM

The artifacts leading to visual discomfort in DIBR-based stereoscopic videos are captured by at least one of the three measures introduced above. We combine the three measures into one 3D vision-based quality measure for stereoscopic DIBR-based videos as follows:

$$\mathbf{3VQM} = K(1 - \mathbf{SO}(\mathbf{SO} \cap \mathbf{TO}))^a(1 - \mathbf{TI})^b(1 - \mathbf{TO})^c, \tag{19}$$

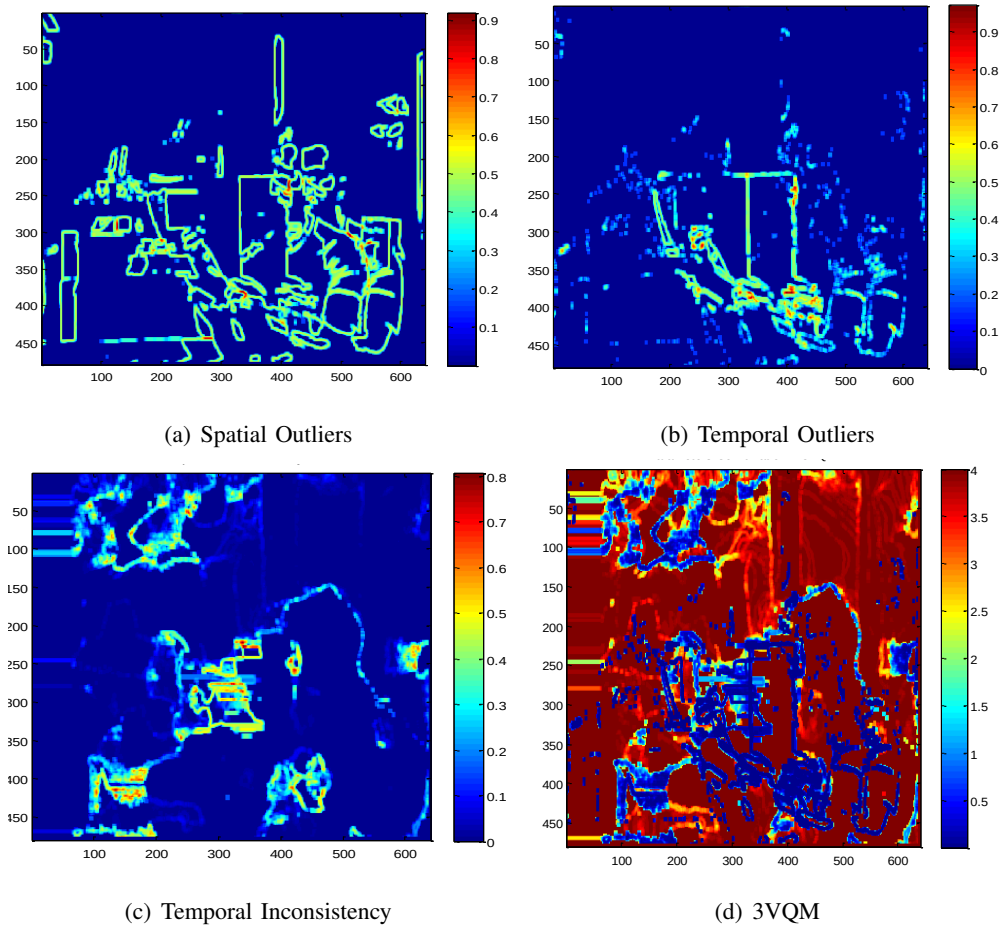
where **SO**, **TO**, and **TI** are normalized to the range 0 to 1 and a, b, and c are constants which were empirically determined by running several training sequences. ($\mathbf{SO} \cap \mathbf{TO}$) is the logical intersection of **SO** and **TO** included in the equation to avoid accounting the outlier distortion more than once [2]. K is a constant for scaling where **3VQM** ranges from 0 for lowest quality to K for highest quality. The overall quality measure is calculated as the mean of the values in the matrix **3VQM**. If **3VQM** is calculated on *ideal depth* derived from a full-reference case it will be referred to as **FR-3VQM**, otherwise if it is derived from no-reference case it will be referred to as **NR-3VQM**. The **FR-3VQM** map of the frame in Fig. 8 is shown in Fig. 9(d).

Fig. 8. A single frame chosen from a right view video generated through DIBR. The depth maps were obtained using stereo matching.

VII. Experimental Results

In order to test the performance of *3VQM*, we conducted an extensive subjective quality assessment study. First we produced a database of DIBR generated video sequences. The original video sequences used are a combination of MPEG sequences [35] and sequences captured using Point Grey's Bumblebee2

[2]For numerical values all nonzero values in the $\cap$ are considered as 1's



(a) Spatial Outliers

(b) Temporal Outliers

(c) Temporal Inconsistency

(d) 3VQM

Fig. 9. Distortion measures for the frame shown in Fig. 8.

camera. To simulate different types of color and depth video distortions , the sequences were processed by three different applications: depth and colored video H.264 based compression, depth estimation (stereo matching), and depth from 2D to 3D conversion using color information [39]. The experiments were conducted using a Samsung 2233RZ display with the shutter glass solution from NVIDIA. The testing conditions were chosen to be consistent with the new requirements for subjective subjective video quality assessment methodologies for 3DTV described in [12]. In these experiments, we recruited 24 volunteers who were mostly engineers with little to no previous experience of 3D video processing. Each volunteer was asked to assign each video sequence with a score indicating his/her assessment of the quality of that video. The subjects were not screened for color blindness or vision problems, and their verbal expression of the soundness of their (corrected) vision was considered sufficient. The quality was defined as the extent to which the distortions were visible and annoying. The raw scores for each subject were collected

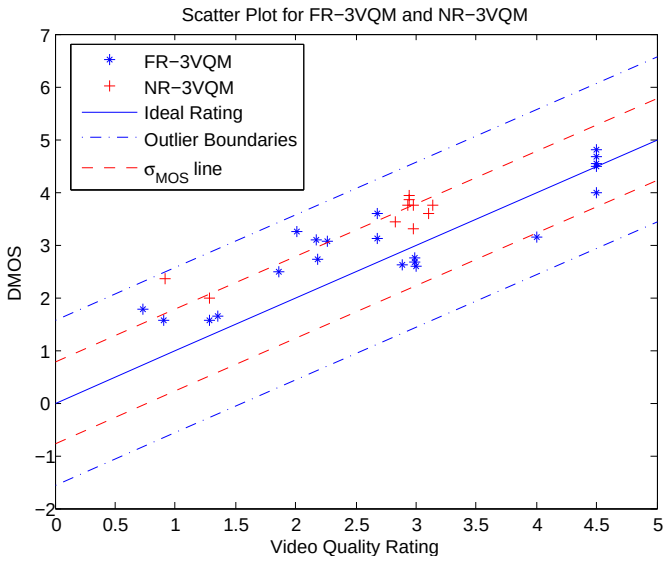
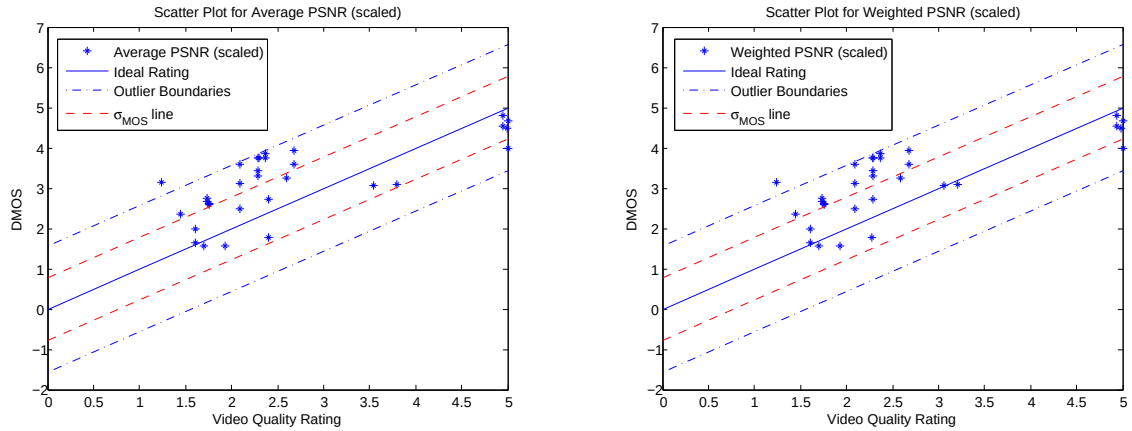


Fig. 10. Scatter plots for both the full-reference measure 3VQM and the no-reference measure NR−3VQM ($d = 5$).

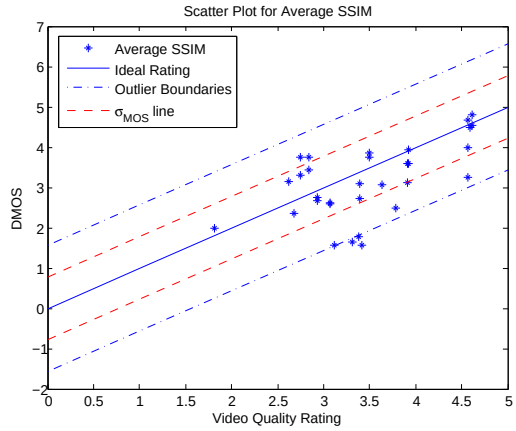
and processed to give Mean Opinion Scores (MOS) and a Difference Mean Opinion Score (DMOS) for each distorted videos. The tested videos were a total of 31 video sequences each of 30 seconds in length. The DMOS results for the video sequences were divided into two groups. For the 21 video sequences of first group we had both the reference distortion-free video the original depth (before processing) and hence the objective quality was measured using the full-reference. However, for the 10 video sequences of the second group we had no information regarding the original depth or the reference distortion-free videos. As a result, the objective quality of the second group was measured using the no-reference measure. Fig. 10 shows the scatter plot for both the (**FR-3VQM**) and the (**NR-3VQM**) measures versus DMOS. To give the values of the **3VQM** a meaningful representation as well as making it easier to compare against the MOS values we have set $K = 5$ in (19). The constants a, b and c were determined after a training experiment conducted using three video sequences in which three different volunteers were asked to rate the synthesized videos. The synthesized videos that used in the training experiment were not used in the subjective experiment and the volunteers who evaluated the training sequence were not asked to perform the subjective experiments so that we can make sure that will not bias our results. Consequently, the constants were set to the following values: $a = 8$, $b = 8$, and $c = 6$.

The results in Fig. 10 show that both **FR-3VQM** and **NR-3VQM** objective ratings are inside the outlier boundary defined by the quality ratings that are greater that two DMOS standard deviation away from the ideal rating. We also notice that more that 80% of the objective ratings fall inside the one σ_{DMOS}

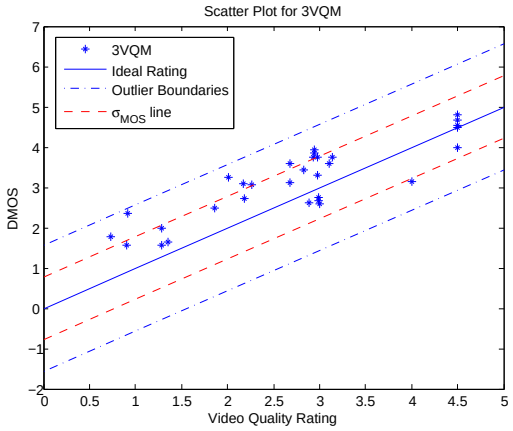


(a) *Average PSNR*

(b) *Weighted Average PSNR*



(c) *Average SSIM*



(d) *3VQM*

Fig. 11. Scatter plots for the four objective quality criteria: *Average PSNR*, *Weighted Average PSNR*,*Average SSIM*, and *3VQM*. The Image Quality Ratings were all scaled to the *MOS* range $[0, 5]$ for comparison. The dashed lines with dots (-.-.) in blue indicate the outliers' boundary and straight line in blue (middle) indicate the ideal image quality rating. A point is considered an outlier if the distance from the ideal is greater than twice the *DMOS* standard deviation [38]. The *DMOS* standard deviation line is shown in dashed red (—-).

boundary. This means that the **3VQM** measure is significantly consistent with the subjective scores and has no outliers.

We compared the performance of `3VQM` against three quality measures. The first quality measure is the average PSNR which is calculated as the average PSNR of the left and right view. The second quality measure is the weighted average PSNR proposed in [21]. Finally, the third quality measure is the average structural similarity (SSIM) of the left and right image [40]. The scatter plots of *DMOS* versus the image

quality ratings for the four objective quality measures (*average PSNR*, *weighted average PSNR*, *average SSIM* and *3VQM*) are shown in Fig. 11. A point is considered an outlier if the distance from the ideal is greater than twice the *DMOS* standard deviation [38]. The plots show that while *3VQM* has zero outlier points, all other measures have points outside the outlier boundaries. The percentage of outlier points in a quality measure is an indicator for consistency. The results are a proof that *3VQM* ratings have no outlier points and hence are significantly more consistent than the other quality measures. The plots in Fig. 11 also show that the *3VQM* values are distributed almost evenly from bad to excellent (1 through 5) thus indicating a coherency.

Table I shows the validations scores for the objective quality measures. Following the *VQEG* recommendations in [38], the validation scores that are used in this paper are the root mean squared error (RMSE), the Pearson linear correlation coefficient (CC), the Spearman rank order correlation coefficient (ROCC), the mean absolute error (MAE), and the Outlier Ratio (OR). These validation scores express the relationships between each quality measure and the subjective ratings. Higher CC and ROCC values indicate increased coherency for the objective quality measure predictions. ROCC is also a metric used to evaluate the monotonicity of the objective quality measure predictions. The RMSE and MAE on the other hand are measures of accuracy of the predictions, where lower RMSE and MAE values mean a more accurate predictions. Moreover, the Outlier Ratio (OR) is a measure of consistency where values closer to zero indicate better consistency in the quality measure predictions.

In the table we also compare the validation scores of the `NR-3VQM` as we increase the block size d. The results show that as the value of d increases the root means quare error (`RMSE`) of the subjective results and the no-reference measure increase as well. Moreover, as the block size increases, the percentage of outliers increases. Nevertheless, from our experiments, a block size of $d = 2$ or $d = 5$ has a low `RMSE`, high correlation values and no outliers for all the videos in our experiment. The reason is that as we increase the block size d, the horizontal shift applied to block in the reference view $\bar{I}_r$ will less likely correspond to the right block in the rendered virtual view $\bar{I}_v$. The results in Table I also show that the no-reference measure value for $d = 2$ has lower `RMSE` and `MAE` values but slightly lower correlation values. This indicates that for $d = 2$ the no-reference measure has higher accuracy but slightly less coherency than full-reference. With $d = 5$ the `RMSE` and `MAE` values are higher however both Pearson linear correlation coefficient (CC) and Spearman rank order correlation coefficient (ROCC) values improved. We can see that the no-reference measure with $d = 5$ is more coherent and is closer in performance to the full-reference than other cases. Outlier ratio is zero except at $d = 100$ which is due to the fact that the correlation is eliminated at large block size. For small block sizes the outlier ratio indicates a very

consistent quality predictions for the no-reference measure for small block size values. The validation scores in Table I for the combination of the full-reference and the no-reference measures show that *ideal depth* evaluation for visual discomfort yields a very accurate, coherent and consistent objective quality prediction for DIBR-based stereoscopic videos.

	RMSE	CC	ROCC	MAE	OR	σ_{DMOS}
Average PSNR	0.9464	0.7311	0.7149	0.822	0.1944	0.7885
Weighted Average PSNR	0.9354	0.7546	0.7766	0.7899	0.1944	0.7885
Average SSIM	0.8062	0.5979	0.542	0.6213	0.1299	0.7885
FR-3VQM	0.6158	0.8942	0.7890	0.5173	0	1.0082
NR-3VQM ($d = 2$)	0.5870	0.8529	0.1180	0.5094	0	0.6652
NR-3VQM ($d = 5$)	0.6384	0.8662	0.4445	0.5551	0	0.6652
NR-3VQM ($d = 10$)	0.7139	0.8762	0.1180	0.6440	0	0.6652
NR-3VQM ($d = 100$)	1.6857	0.9157	0.1003	1.6632	0.8	0.6652
FR-3VQM and NR-3VQM ($d = 5$)	0.6875	0.8728	0.7894	0.5967	0	0.7885

TABLE I

VALIDATION SCORES FOR THE FULL-REFERENCE, THE NO-REFERENCE AND THE COMBINATION OF BOTH THE FULL-REFERENCE AND THE NO-REFERENCE MEASURES. THE VALIDATION CRITERIA ARE: ROOT MEAN SQUARED ERROR(RMS), PEARSON LINEAR CORRELATION COEFFICIENT (CC), SPEARMAN RANK ORDER CORRELATION COEFFICIENT (ROCC), MEAN ABSOLUTE ERROR (MAE), OUTLIER RATIO (OR) AND THE STANDARD DEVIATION OF THE DMOS VALUES σ_{DMOS}.

The results in the Table I shows that *3VQM* values represented by **FR-3VQM** and **NR-3VQM** ($d = 5$) has the least RMSE and MAE values among the computed objective quality measures. In addition, the RMSE for *3VQN* is less than one standard deviation of the DMOS values ($\delta_{DMOS} = 0.7885$), which actually is an indication that *3VQM* is relatively an accurate prediction of the quality for DIBR-based 3D videos. The Pearson linear correlation (CC) and Spearman rank order correlation coefficient (ROCC) values for *3VQM* also outperforms the three other quality measures. Higher CC values mean that *3VQM* is more coherent than *average PSNR*, *weighted average PSNR* and *average SSIM*. Higher ROCC values also indicate a significant gain in monotonicity of quality predictions using *3VQM* over the closest quality measure *average SSIM*. The results also show that *3VQM* has a zero outlier ratio (OR) and therefore is the most consistent quality measure among all the ones above.

Overall, *FR-3VQM* is the most accurate, coherent and consistent among the objective measures represented in this paper. The results also show that *average SSIM* has a lower OR value than *average PSNR*

and *weighted average PSNR*, which indicates that *average SSIM* is more consistent. *average SSIM* is second to *3VQM* in accuracy (RMSE and MAE). However, *average SSIM* has the least coherency (CC and ROCC).

VIII. Conclusion

In this paper, we presented a new family of methods to objectively evaluate the quality of stereoscopic 3D videos generated by DIBR. This new family rely on a new concept we introduced;i.e. ideal depth. First we showed how to derive an ideal depth estimate at each pixel value that would constitute a distortion-free rendered video. We showed how to derive the ideal depth for a full-reference and no-reference cases. The ideal depth estimate was then used to derive three distortion measures to objectify the visual discomfort in the stereoscopic videos. The three measures are temporal outliers (`TO`), temporal inconsistencies (`TI`), and spatial outliers (`SO`). The combination of the three measures constituted a vision-based quality measure for 3D DIBR-based videos, `3VQM`. Finally, `3VQM` was verified against a fully conducted subjective evaluation and compared to three other quality measures. The results show that our proposed measure is significantly accurate, coherent and consistent with the subjective scores. The results have also shown that the predictions of the no-reference measure (**NR-3VQM**) highly correlates with subjective scores and is fairly close in performance to the full-reference **FR-3VQM**.

IX. Acknowledgements

This work has been conducted in part during an internship in the office of chief scientist in Qualcomm, San Diego.

References

[1] A. Kubota, A. Smolic, M. Magnor, M. Tanimoto, T. Chen, and C. Zhang, "Multiview Imaging and 3DTV," *IEEE Signal Proc. Magazine*, vol. 24, no. 6, pp. 10–21, Nov 2007.

[2] S.C. Chan, Heung-Yeung Shum, and King-To Ng, "Image-based rendering and synthesis," *Signal Processing Magazine, IEEE*, vol. 24, no. 6, pp. 22 –33, nov. 2007.

[3] C. Fehn, "Depth-image-based Rendering (DIBR), Compression, And Transmission For A New Approach On 3DTV," *Proc. of SPIE*, vol. 5291, pp. 93–104, 2004.

[4] W.H.A. Bruls, C. Varekamp, R.K. Gunnewiek, B. Barenbrug, and A. Bourge, "Enabling Introduction of Stereoscopic (3D) Video: Formats and Compression Standards," in *ICIP 2007*, 16 2007-oct. 19 2007, vol. 1, pp. I –89 –I –92.

[5] L. McMillan, *An Image Based Approach to Three-Dimensional Computer Graphics*, Ph.D. thesis, Univ. of North Carolina at Chapell Hill, NC,USA, 1997.

[6] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth Map Creation and Image-based Rendering for Advanced 3DTV Services Providing Interoperability and Scalability," *Image Commun.*, vol. 22, no. 2, pp. 217–234, 2007.

[7] W. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth Map Distortion Analysis for View Rendering and Depth Coding," in *ICIP 2009*, 2009, pp. 721–724.

[8] M. Lambooij, W. IJsselsteijn, and I. Heynderickx, "Stereoscopic Displays And Visual Comfort: A Review," *SPIE Jour. of Image Science Tech.*, June 2009.

[9] P. Aflaki, M.M. Hannuksela, J. Ha andkkinen, P. Lindroos, and M. Gabbouj, "Subjective study on compressed asymmetric stereoscopic video," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, sept. 2010, pp. 4021 –4024.

[10] Francesca De Simone Lutz Goldmann and Touradj Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," *Proc. SPIE*, vol. 7526, pp. 75260S, 2010.

[11] Pieter Seuntiens, Lydia Meesters, and Wijnand Ijsselsteijn, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric jpeg coding and camera separation," *ACM Trans. Appl. Percept.*, vol. 3, pp. 95–109, April 2006.

[12] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "New Requirements Of Subjective Video Quality Assessment Methodologies For 3DTV," in *VPQM*, Scottsdale,US, 2010.

[13] VQEG, "Proposal for 3D evaluation test plan in VQEG.," in *Video Quality Experts Group meetings, Japan, http://www.vqeg.org/*, June 2011.

[14] Hang Shao, Xun Cao, and Guihua Er, "Objective Quality Assessment of Depth Image Based Rendering in 3DTV System," in *Proc. 3DTV Conf.*, 2009, pp. 1–4.

[15] A. Tikanmaki, A. Gotchev, A. Smolic, and K. Miller, "Quality Assessment of 3D Video in Rate Allocation Experiments," in *Proc. IEEE Int. Symp. Consumer Electronics ISCE 2008*, 2008, pp. 1–4.

[16] C. T. E. R. Hewage, S. T. Worrall, S. Dogan, S. Villette, and A. M. Kondoz, "Quality Evaluation of Color Plus Depth Map-Based Stereoscopic Video," vol. 3, no. 2, pp. 304–318, 2009.

[17] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Using disparity for quality assessment of stereoscopic images," in *Proc. 15th IEEE Int. Conf. Image Processing ICIP 2008*, 2008, pp. 389–392.

[18] A. Tikanmaki, A. Gotchev, A. Smolic, and K. Miller, "Quality assessment of 3D video in rate allocation experiments," in *IEEE symposium on Consumer Electronics*, April 2008, pp. 1–4.

[19] Jiangbo Lu, Qiong Yang, and G. Lafruit, "Interpolation Error As A Quality Metric For Stereo: Robust, Or Not?," in *ICASSP 2009*, 2009, pp. 977–980.

[20] Z. M. P. Sazzad, S. Yamanaka, and Y. Horita, "Spatio-temporal Segmentation Based Continuous No-reference Stereoscopic Video Quality Prediction," in *QoMEX 2010*, 2010, pp. 106–111.

[21] A.Tekalp N. Ozbek and E. Tunali, "Rate Allocation Between Views in Scalable Stereo Video Coding using an Objective Stereo Video Quality Measure," in *ICASSP*, April 2007, pp. 1045–1048.

[22] J. Starch, J. Kilner, and A. Hilton, "Objective quality assessment in free-viewpoint video production," in *3DTV Conference 2008*, may 2008, pp. 225 –228.

[23] P. Campisi, A. Benoit, P. Callet, and R. Cousseau, "Quality Assessment of Stereoscopic Images," in *(EUSIPCO)*. EURASIP, September 2007.

[24] C. Hewage, S. Worrall, S. Dogan, and A. Kondoz, "Prediction of Stereoscopic Video Quality Using Objective Quality Models of 2-D Video," *IEEE Elect. Letters*, vol. 44, pp. 963–965, July 2008.

[25] Yan Zhang, Ping An, Yanfei Wu, and Zhaoyang Zhang, "A multiview video quality assessment method based on disparity and ssim," in *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*, oct. 2010, pp. 1044 –1047.

[26] A. Mittal, A.K. Moorthy, J. Ghosh, and A.C. Bovik, "Algorithmic assessment of 3d quality of experience for images and videos," in *Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), 2011 IEEE*, jan. 2011, pp. 338 –343.

[27] M. Lambooij, W. IJsselsteijn, D.G. Bouwhuis, and I. Heynderickx, "Evaluation of stereoscopic images: Beyond 2d quality," *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 432 –444, june 2011.

[28] Donghyun Kim and Kwanghoon Sohn, "Visual fatigue prediction for stereoscopic image," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 2, pp. 231 –236, feb. 2011.

[29] L. Zhang and W.J. Tam, "Stereoscopic Image Generation Based On Depth Images For 3DTV," *IEEE Trans. on Broadcasting*, vol. 51, no. 2, pp. 191–199, June 2005.

[30] Quang H. Nguyen, Minh N. Do, and Sanjay J. Patel, "Depth Image-based Rendering From Multiple Cameras with 3D Propagation Algorithm," in *IMMERSCOM '09*, 2009, pp. 1–6.

[31] Sang-Beom Lee and Yo-Sung Ho, "Discontinuity-Adaptive Depth Map Filtering for 3D View Generation," in *IMMERSCOM '09*, 2009, pp. 1–6.

[32] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski, "Layered Depth Images," in *SIGGRAPH '98*, New York, NY, USA, 1998, pp. 231–242.

[33] M. Solh and G. AlRegib, "Hierarchical Hole-Filling (HHF): Depth Image Based Rendering without Depth Map Filtering for 3D-TV," in *IEEE MMSP'10*, Saint-Malo, France, October 4-6 2010.

[34] Qingxiong Yang, Kar-Han Tan, B. Culbertson, and J. Apostolopoulos, "Fusion of active and passive sensors for fast 3d capture," in *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on*, oct. 2010, pp. 69 –74.

[35] "Mobile 3dtv - 3d video database, http://sp.cs.tut.fi/mobile3dtv/stereo-video/," .

[36] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski, "High-quality video view interpolation using a layered representation," *ACM Trans. Graph.*, vol. 23, pp. 600–608, August 2004.

[37] "Middlebury stereo evaluation - version 2, http://vision.middlebury.edu/stereo/eval/," .

[38] VQEG, "Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment.," in *http://www.vqeg.org/*, March 2000.

[39] Y. Chen, R. Zhang, and M. Karczewicz, "Low-complexity 2d to 3d video conversion," in *Stereoscopic Displays and Applications XXII*, February 2011, vol. 7863, p. 78631I.

[40] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. on Image Proc.*, vol. 13, pp. 600–612, 2004.

PLACE
PHOTO
HERE

Mashhour Solh is a computational photograph system engineer with the camera system technology team in the OMAP platform buisness unit at Texas Instruments. Dr. Solh earned a Ph.D. in electrical and computer engineering (ECE) from Georgia Institute of Technology in 2011. He received a degree M.S. degree in ECE from Florida Institute of Technology in 2006, and a B.E. in computer and communications engineering from American University of Beirut in 2005. His research scope includes computational photography, image and video coding, multimedia quality assessment, multi-camera imaging, multimedia signal processing (image, video and audio), 3DTV techniques, signal processing for infrasound signals and Digital Signal Processors.

Dr. Solh has served as web chair of IMMERSCOM 2007, IMMERSCOM 2009 and IMMERSCOM 2011. He is a member of Video Quality Experts Group (VQEG) standardization efforts. Dr. Solh was also vice president of IEEE student branch of Georgia Institute of Technology for 2009, chair of ECE seminar student committee at Georgia Tech 2007-2009, and a member of Georgia Techs student advisory committee 2009-2010. He is a member of IEEE signal processing society and IEEE communications society.

Dr. Solh has received the ECE Graduate Research Assistant Excellence Award (2012), the Franking Antonio Scholarship Award for his outstanding performance during a internship in Qualcomm in summer of 2010, and the travel grant award for the IEEE International Conference on Multimedia and Expo (ICME 2011). Dr. Solh has been awarded the outstanding achievement award for his service as a Florida Tech diplomat (2006). He also earned the Deans Honor List Award several times and the pepsi cola international scholarship while being a student at the American University of Beirut (2001-2005).

PLACE
PHOTO
HERE

Ghassan AlRegib received the Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology in 2003. He joined the Georgia Tech faculty in 2003 and is currently an associate professor in the School of Electrical and Computer Engineering (ECE). His research group is working on projects related to multimedia processing and communications, immersive communications, distributed processing, collaborative systems, quality of experience, and wireless sensor networks.

As a graduate student at Georgia Tech, Dr. AlRegib received the ECE Outstanding Graduate Teaching Award in spring 2001 and both the Center for Signal and Image Processing (CSIP) Research Award and the CSIP Service Award in spring 2003. In 2008, he received the ECE Outstanding Junior Faculty Member Award at Georgia Tech.

Dr. AlRegib was the general co-chair and co-founder of the First International Conference on Immersive Telecommunication (IMMERSCOM) that was held in November 2007. He is currently the steering committee co-chair for IMMERSCOM and was the chair of the Special Sessions Program at the IEEE International Conference on Image Processing (ICIP) in 2006. Dr. AlRegib has also served as an associate editor of the IEEE Signal Processing Magazine, and in 2008, he became the area editor for the IEEE Signal Processing Magazine and the editor-in-chief of the ICST Transactions on Immersive Communications. Dr. AlRegib has served as a session chair and technical program committee member for several international conferences and workshops. He has authored over 60 journal and conference technical papers and has been issued four U.S. patents. He has conducted several consulting jobs for several companies and organizations.

PLACE
PHOTO
HERE

Judit Martinez Bauza received the Ph.D. degree (with Honors) in telecommunications engineering from the Polytechnic University of Catalonia, Spain, in 1998. She has been a researcher at the Industrial Robotics Institute (IRII) of the Spanish High Council for Scientific Research and member of the research staff at the Computer Vision Center, an R&D center founded by the Autonomous University of Barcelona and the Autonomous Government of Catalonia. She has been principal researcher of several industrial and research projects related to computer vision and multimedia technologies. She joined Qualcomm Inc. in 2007 as a staff engineer in the Office of the Chief Scientist and recently moved to the Multimedia R&D group. She has authored several conference and journal papers, submitted and been issued a number of patents, participated in review committees and standardization groups, and has been a member of IEEE societies since 1991. Her research interest encompasses several areas of image/video processing and communication technologies, including all aspects of 3D video, multi-camera imaging, computer vision technologies, efficient algorithms for multimedia systems, multi-resolution mathematical models, inverse problems, compression and imaging systems.