No-Reference Quality Metric for Depth Maps

S. Milani, D. Ferrario, S. Tubaro

Dipartimento di Elettronica ed Informazione, Politecnico di Milano, Milano, Italy

e-mail: milani@elet.polimi.it,daniele.ferrario86@gmail.com,stefano.tubaro@polimi.it

***Abstract*—The performances of several 3D imaging/video applications (going from 3DTV to video surveillance) benefit from the estimation or acquisition of accurate and high quality depth maps. However, the characteristics of depth information is strongly affected by the procedure employed in its acquisition or estimation (e.g., stereo evaluation, ToF cameras, structured light sensors, etc.), and the very definition of "quality" for a depth map is still under investigation.**

In this paper we proposed an unsupervised quality metric for depth information in Depth Image Based Rendering signals that predicts the accuracy in synthesizing 3D models and lateral views by using the considered depth information. The metric has been tested on depth maps generate with different algorithms and sensors. Moreover, experimental results show how it is possible to progressively improve the performance of 3D modelization by controlling the device/algorithm with this metric.

***Index Terms*—no-reference metric, quality evaluation, depth maps, MS Kinect, DIBR.**

I. Introduction

Geometry information has recently proved to be crucial in many imaging and control applications. As a consequence, both academic and industrial communities have entailed a significant research effort towards the design of new sensors and algorithms for geometry estimation. Time-of-Flight cameras [1], structured light 3D scanners [2], stereoscopic and multicamera systems, as examples, permit nowadays a real-time acquisition of both static and dynamic elements with different accuracies and resolutions.

Although these systems permits a quick depth estimation, severe noise levels (due to illumination conditions, characteristics of the device/algorithm, time constraints) affects the accuracy of depth estimate and the morphology of objects. The irregularity of object edges in depth maps (see Fig. 1) lead to highly-distorted shapes and wrong repositioning of 3D points whenever a synthesized view has to be generated. These effects are mainly evident in infrared (IR) depth cameras due to the presence of external IR noise (solar light) and to the triangulation strategy they adopt for depth estimation [2]. In addition, many artifacts for depth signals also appear in stereoscopic or multicamera systems. Consequently, we have to take into consideration some operations performed after depth map acquisition (e.g., compression), which may alter the quality of the displayed signal. From these premises, quality evaluation strategies for depth maps are nowadays significant in many applications [3] since they permits understanding whether the acquired data are valid or not.

In this paper, we present a no-reference unsupervised metric that mainly targets the morphology of the objects in the estimated depth map. The proposed strategy operated on DIBR video + depth signals, i.e., depth maps with an associated color view, and tries to map irregularities along the edges in a quality value. More precisely, it computes segmentation features from color information and from depth maps independently and the checks their consistency. The proposed solution has been tested on depth maps obtained from structured light sensors, Structure-from-Motion (SfM) algorithms, and multicamera systems. Moreover, it has also been employed in real time acquisition in order to maximize the performance of devices.

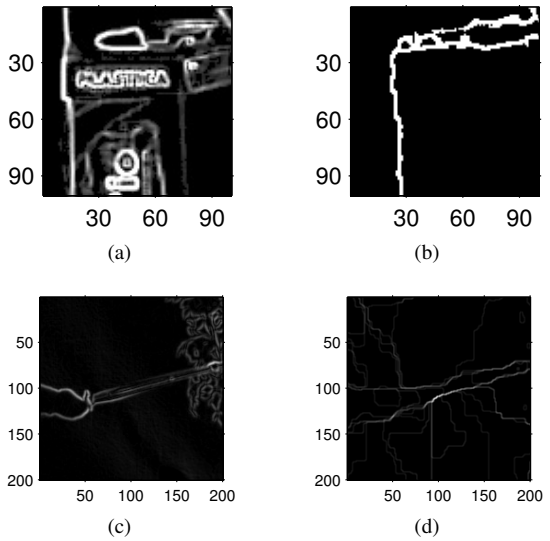


Fig. 1. Edges (detail) computed from Sobelian on different signals acquired with different methods. a) RGB view b) depth map for dataset `bins` (MS Kinect) ; c) color view 3 d) depth map for sequence `kendo` (multicamera).

In the following, Section II presents a short review of objective quality evaluation strategy for depth signals. Section III describes the proposed metric, while Section V reports some experimental results displaying the correlation between the proposed solution and other quality metrics, together with the improvement in performance of some 3D systems including the proposed metric. Section VII draws the final conclusions.

II. State-of-the-art of objective depth quality metrics

A lot of research work has been devoted to the perceptual evaluation of 3D scene using different kinds of signals and displaying devices. In our work, we focus on objective quality metrics.

In DIBR and multiview systems, the quality of 3D signals can be measured by synthesizing some side views and comparing them with some reference information acquired

with different cameras [4]. In stereoscopic systems, left-to-right checking and reliability checking is applied at the end of the estimation process to reduce the error of the estimated depth map [5]. In 3D acquisition and modeling, the precision of depth maps can be measured comparing the estimated data with respect to the real measures of the object. In all these cases, quality measurement requires some reference data to be compared with the generated model. This assumption may not be met in many practical situations or the characteristics of the reference data may prevent a reliable evaluation of the accuracy of depth maps. Reference data are not available for the simplest 3D acquisition and transmission systems, e.g., ToF cameras or structured light systems like MS Kinect or DIBR video + depth signals. In these cases, depth information is associated to a single RGB view of the scene. In other cases, some reference data or view may be available but, because of their quality or bad calibration, they do not provide a reliable benchmark for depth evaluation. For some multicamera systems, different sensors may present different illumination and aberration characteristics which would bias the results of comparing a reference view with its warped version. The same conclusions can be drawn for signals that have been compressed [3]. Passive stereoscopic systems compute depth information by performing a local optimization: the obtained depth map optimizes warping operations between couples of views, but its quality needs to be measured with respect to some additional reference views. Moreover, computing the average distortion between views and their warped versions does not allow to accurately evaluate the effects of noise along the morphologies of the objects. In addition to these problems, we must also take into consideration the computational complexity required by some of the proposed evaluation methods and the impossibility of performing quality evaluation in real time during the acquisition of the model.

From these premises, we have designed a quality evaluation strategy that pays attention to morphology of objects, does not need extra references, and has limited complexity requirements.

III. The proposed no-reference metric

As mentioned in Section I, quality evaluation for depth maps must take into consideration the alteration on morphology comparing object shapes with some low-noise references. These references are obtained from an associated color view of the scene, where edges proves to be more regular as many works on interpolation and denoising have proved [6], [7]. At first, the proposed metric computes a set of morphological features from depth maps. Then the computed set of features is compared with the original color view and the compatibility level is used as a measurement of noise for the proposed depth map.

Given the input depth map D (with size $h \times w$) and the associated equally-sized color view I (that can be split into the three RGB color components $I_R(x,y)$, $I_G(x,y)$, and $I_B(x,y)$), the algorithm segments the depth map first. We call the segmentation map S_D. Then, the compatibility of segmentation is checked using an unsupervised no-reference quality metric for the segmentation and the color signal. The following subsections will describe this process in detail.

A. Segmentation

In our implementation, we used a simple segmentation algorithm operated on depth values $D(x,y)$. More precisely, depth values $D(x,y)$ are clustered into N classes using a simple k-means algorithm.From this operations, the segmentation S_D is computed, assigning to the pixel at position (x,y) the index i of the cluster it belongs to.

The parameter N is chosen by performing an iterative optimization of the segmentation using the no-reference segmentation quality metric F_{RC} [8], which will be described in the following section.

Additional tests were performed using more complex segmentation strategies (like that reported in [9]). Although the precision in the segmentation is much better, the impact on the final performance of the metric is limited.

B. The no-reference segmentation quality metric F_{RC}

In the literature, several no-reference quality metrics have been proposed at different times to evaluate the accuracy of segmentation. According to the review in [10], the F_{RC} solution [8] provides good performances and precision in characterizing the quality of segmentation. In our approach, we design a color-based and a depth-based versions of the metric in order to validate both color and depth images.

A segmentation $S(x,y)$ can be seen as a bi-dimensional function that maps coordinates (x,y) of image pixels (with $(x,y) \in P \subset \mathbb{Z}^2$) to the index of the corresponding segment. Given the set of N segments $\mathcal{S} = \{s_i \subset P : (x,y) \in s_i \text{ and } S(x,y) = i\}$, F_{RC} combines two different sets of disparity features, which are function of the segmentation map S and of the input image I.

The intra-region disparity $d(s_i)$ is measured as

$$d_{intra}(I,S) = \frac{1}{N}\sum_{i=1}^{N}\frac{|s_i|}{h \cdot w} e(I,s_i) \tag{1}$$

where

$$e(I,s_i) = \frac{1}{3}\sum_{c=R,G,B}\sum_{(x,y)\in s_i} \|I_c(x,y) - \hat{I}_c(s_i)\|^2. \tag{2}$$

The value $\hat{I}_c(s_i)$ denotes the average of the values $I_c(x,y)$ for the pixels in s_i, while $|\cdot|$ denotes the cardinality of the set.

The inter-region disparity is, in our case, the average

$$d_{inter}(I,S) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|S_n(s_i)|}\sum_{s_j \in S_n(s_i)}\frac{\|B_i - d_j\|_2}{\|B_i\|_2 + \|B_j\|_2} \tag{3}$$

where $B_i = [\hat{I}_R(s_i)\hat{I}_G(s_i)\hat{I}_B(s_i)]$ is the baricenter of region s_i and $S_n(s_i)$ is the set of neighboring segments for s_i. The final quality metric can be represented

$$F_{RC}(I,S) = \frac{d_{inter}(I,S) - d_{intra}(I,S)}{2}. \tag{4}$$

It is possible to compute a similar measure $F'_{RC}(D,S)$ for depth maps and their segmentation as well. In this case, only one component (D) is present.

IV. The proposed no-reference quality metric for depth maps

After the computation of the segmentation S_D, it is possible to evaluate the quality of segmentations using F_{RC} More precisely, the metric F_{RC} check the quality of segmentation by evaluating the consistency of the originating signal with respect to the computed segments.

In our approach, segmentation is refined by iterating the segmentation routine over different values of the parameter N and computing the metric $F'_{RC}(D,S_D)$. After the iteration, the segmentation process chooses the map S_D that maximizes $F'_{RC}(D,S_D)$.

In order to check the consistencies of the segmentation with respect to the color signal, we compute the F_{RC} for the segmentation S_D using the signal I in place of D ($F_{RC}(I,S_D)$).

Experimental results proved that these values are highly correlated with the quality (in terms of distortion) of the 3D models we are considering. More precisely, we correlated these cross-computed F_{RC} values with the MSE of warped view with respect to its reference and the Euclidean distance between points of the 3D model under evaluation with respect to a reference tridimensional model of the scene.

In both cases, distortion measures have proved to be correlated with the metric

$$Q_I = \frac{k_1}{F_{RC}(I,S_D)} + k_2\ F_{RC}(I,S_D) + k_3, \qquad (5)$$

where the parameters k_1, k_2, k_3 depends on the specific application we are considering. More precisely, their values are found performing a data regression on a training set of reference samples. The effectiveness of their value has been evaluated on a different set of acquisitions (test set).

Reversing the estimation procedure, we tested the correlation between the distortion and the opposite metric obtained by segmenting I and computing the segmentation consistency on D. The parameter $Q_D = k_1/F_{RC}(I,S_D) + k_2\ F_{RC}(I,S_D) + k_3$ has proved to have a minor correlation with respect to the measured distortion.

The following sections will present how Q_I models the precision of the 3D model very well and how it can be applied to modulate the number of acquisition required for an infrared depth sensor in order to acquired a 3D model with a given accuracy level.

V. Experimental results with respect to other metrics

In order to validate the proposed metric, we considered different DIBR scenarios where depth signals are corrupted by different types of noise. In these tests, our aim was to show the correlation existing between the proposed solution and different referenced quality metrics.

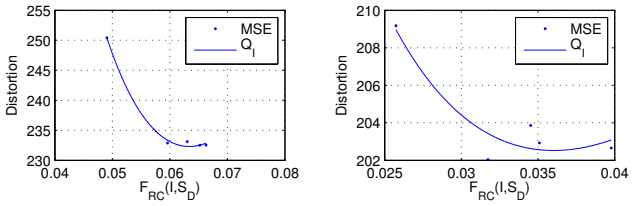


Fig. 2. MSE vs. $F_{RC}(I,S_D)$ values for quantization noise in multiview systems. (a) breakdancers (b) ballet.

A. DIBR signals from multicamera systems

In the first scenario we applied our metric to DIBR signals acquired via a multicamera system, where depth maps are computed via disparity estimation algorithms. We considered the depth maps generated for the central camera in the datasets . The referenced quality metric is the MSE between one the side views (the left one) and its warped version obtained from the color and the depth information of the central camera. In this evaluation we corrupted depth information with different amount of coding noise. More precisely, we coded the depth map for datasets `breakdancers` and `ballet` at different compression qualities (in this case we adopted a simple Intra coding). Figure 2 reports the MSE and the value of Q_I for different compression ratios. It is possible to notice that the correlation between the proposed metric and the actual MSE is quite good. As a matter of fact, the proposed method proves to be quite effective in performing a non-referenced evaluation of the quality of transmitted DIBR signals.

B. DIBR signals from infrared sensors

In the previous subsection we have shown that the proposed metric accurately models the distortion associated to warping in multicamera systems. However, the proposed metric turns to be crucial in those cases where having some side views to evaluate the accuracy of depth is not possible or computationally-demanding. This is the case of real time depth estimation performed by infrared sensors like MS Kinect [2] or ToF cameras [1]. From these premises, in a second evaluation scenario, we considered a video+depth signal obtained via an infrared structured light sensor (in our case, the MS Xbox Kinect). In this case, the main source of noise in depth estimation is infrared radiation coming from external sources (e.g., sunlight, reflections).

The accuracy of the proposed metric was tested using two different reference metrics. A first metric is the MSE between the image acquired by an additional side camera and the warped version generated from the 3D model estimated by the Kinect sensor.

Figure 3 reports the results obtained for different datasets obtained under different illumination conditions. Note that two curves are plotted in order to distinguish denser point clouds. In fact, two curves are reported: the first related to datasets with more than 50 % of samples with a valid depth value (because of illumination noise), and the second related to

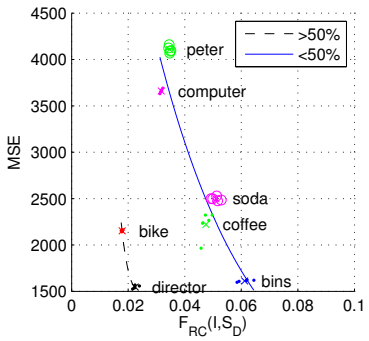
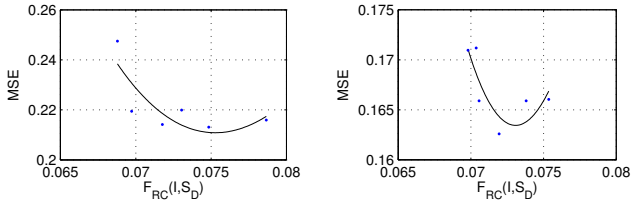


Fig. 3. MSE vs. $F_{RC}(I, S_D)$ for Kinect-acquired scenes where distortion is measured with respect to a warped view.



(a) (b)

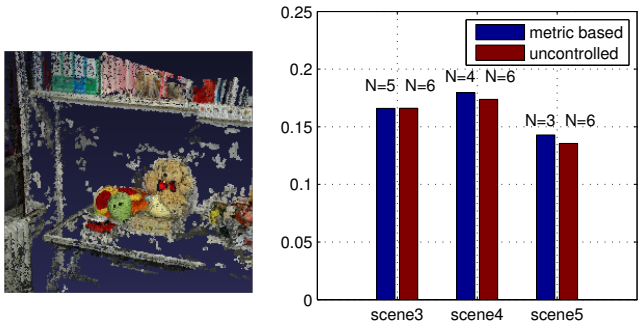
Fig. 4. MSE vs. $F_{RC}(I, S_D)$ for Kinect-acquired scenes where distortion is measured with respect to a point-cloud reference model. The scene was acquired with noise source at different distances d. (a) `scene3` (where $d = 2.5$ m), (b) `scene4` (where $d = 1.5$ m).

dataset with less than 50 % of valid samples.

A second metric considered the MSE between the 3D points of the Kinect-generated point-cloud model and the 3D points of a reference point-cloud model estimated via a Structure-from-Motion (SfM) algorithm (see Fig. 4). Here, we assume that SfM solutions provide a better accuracy in estimating the geometrical location of 3D points since they can rely on a wider set of images for the same object, taken from different viewpoints. As result, the minimization process proves to be more accurate. Points are matched via an affine transformation that was estimated from a reduced set of known corresponding points between the models. The transformation is computed via the ICP algorithm, and the transformed points of the Kinect model are matched with the reference points minimizing the Euclidean distance between geometric and color components. In this evaluation, we adopted the same 3D scene but we varied the noise source (a incandescence lamp positioned at different distances from the scene). It is possible to notice from Figure 4, that the previous behavior of the metric is verified once again.

VI. Experimental results on quality improvement

The previous section has shown how the proposed strategy is strongly correlated with other metric. In this section, we show how the proposed solution can be used to drive the acquisition process for a Kinect sensor without requiring a side view to evaluate the accuracy of the 3D model. The depth map acquired by the Kinect sensor are affected by a significant noise level that is not systematic. Therefore, different acquisitions of the same scene at different times



(a) (b)

Fig. 5. Quality evaluation merging different numbers of acquisitions for Kinect sensor. The number of acquisitions is reported over each bar. The distance d of the source of noise from the scene is varied: `scene3` $d = 2.5$ m; `scene4` $d = 1.5$ m; `scene5` no noise. (a) reference point cloud model (b) MSE and number of acquisitions.

presents different and independent artifacts. These allow to infer that the quality of the 3D model generated via Kinect can be improved combining different acquisitions in order to compensate the artifacts and the missing point of a single view. In this work, we assume that different Kinect acquisition of the same scene/object are averaged in order to refine the depth values and increase the number of valid depth values. This process (frequently exploited in SLAM) can be significantly improved and speeded if the sensor knows whenever it has acquired a sufficient number of depth maps that permit building a sufficiently-accurate model.The proposed metric

Figure 5(b) reports the accuracy of the Kinect point-cloud model (in terms of MSE with respect to a reference model obtained via SfM) after combining a variable number of depth maps together. We report as well the number of acquisitions required to generate the model (superimposed to each bar). The `metric based` approach stops acquiring depth maps as soon as the actual value of $F_{RC}(I, S_D)$ verifies $\delta Q_I / \delta F_{RC}(I, S_D) = 0$. It is possible to notice that the metric-driven strategy, compared with a permits obtaining a better precision with a lower number of acquisition improving the speed of 3D modeling process.

VII. Conclusions

The paper presented a no-reference quality metric for DIBR signals. The proposed metric exploits no-reference segmentation metrics applied to depth-based segmentations. The proposed solution has been successfully tested on different types of video+depth signals that were obtained via different methods or devices. Moreover, it has also been applied to a data acquisition strategy for the MS Kinect sensors that refine the obtained 3D model averaging different sequential acquisitions. The proposed strategy permits minimizing the number of acquisition improving the speed and reducing the required computational load of the approach. Future work will be devoted to analyze the correlation between depth quality perception of human users and the value suggested by the proposed metric.

REFERENCES

[1] S.B. Gokturk, H. Yalcin, and C. Bamji, "A time-of-flight depth sensor - system description, issues and solutions," in *Proc. of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2004)*, June 27 – July 2, 2004, vol. 3, p. 35.

[2] IHS iSuppli, "The teardown: The kinect for xbox 360," *IET Engineering Technology*, vol. 6, no. 3, pp. 94 –95, Apr. 2011.

[3] M. Solh and G. AlRegib, "A no-reference quality measure for dibr-based 3d videos," in *Proc of the 2011 IEEE International Conference on Multimedia and Expo (ICME 2011)*, July11 – 15, 2011, pp. 1–6.

[4] C. Fehn, "3D-TV Using Depth-Image-Based Rendering (DIBR)," in *Proc. of PCS 2004*, San Francisco, CA, USA, Dec. 2004.

[5] Andrea Fusiello and Vito Roberto, "Efficient stereo with multiple windowing," in *In CVPR*. 1997, pp. 858–863, IEEE Computer Society Press.

[6] Valeria Garro, Carlo dal Mutto, Pietro Zanuttigh, and Guido M. Cortelazzo, "A novel interpolation scheme for range data with side information," in *Proceedings of the 2009 Conference for Visual Media Production*, Washington, DC, USA, 2009, CVMP '09, pp. 52–60, IEEE Computer Society.

[7] S. Milani and G. Calvagno, "Joint denoising and interpolation of depth maps for ms kinect sensors," in *Proc. of the* 37^{th} *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Mar. 25 – 30, 2012, pp. 797–800.

[8] C. Rosenberger and K. Chehdi, "Genetic fusion: application to multi-components image segmentation," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, 2000, vol. 6, pp. 2223 –2226 vol.4.

[9] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167 – 181, Sept. 2004.

[10] Hui Zhang, Jason E. Fritts, and Sally A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Comput. Vis. Image Underst.*, vol. 110, no. 2, pp. 260–280, May 2008.