Infrared structured light sensors are widely employed for control applications, gaming, acquisition of dynamic and static 3D scenes. Recent developments have led to the availability on the market of low-cost sensors which prove to be extremely sensitive to noise, light conditions, materials, the surface nature of the objects, and their distance from the camera. As a matter of fact, accurate denoising and interpolation strategies are needed.

The paper presents a quality enhancement strategy for depth maps targeting low-cost IR structured light sensors. The approach has been tested using the MS Xbox Kinect device in both indoor and outdoor scenarios under different light conditions.

Index Terms — interpolation, denoising, MS Kinect, 3D scanning, structured light camera, infrared sensor.

1. INTRODUCTION

The recent availability of low-cost range cameras has shaken the ICT world leading to a flourishing of new object recognition applications, human-computer interfaces, and acquisition systems of dynamic 3D scenes. Time-of-Flight cameras [1], structured light 3D scanners [2], multicamera systems allow easy real-time acquisition of dynamic 3D scenes with both static and dynamic elements.

Among these, the Xbox Kinect sensor [2], which includes a standard RGB camera together with an infrared (IR) structured light scanner, has recently proved to be one of the most widely-used sensors thanks to its versatility and the limited cost (see Fig. 2). Unfortunately, despite the strong versatility and the wide range of new applications that these IR devices enable, the resulting depth signal is affected by a significant amount of noise. One of the main reasons for this inconvenience is that for most of the acquired scene there is no control over illumination, and therefore, IR sensors receive a significant amount of radiation that has not been provided by the 3D device itself. Depth sensors also present shot noise related to the radiation, A/D conversion quantization noise, and thermal noise. Moreover, the artifacts along object boundaries and the limited resolution of the acquired depth maps utterly increase the need for interpolating and denoising algorithms.

Several works have proposed novel denoising algorithms to improve the quality of the acquired depth maps. One of these considers the confidence values to denoise depth acquired via a ToF camera, while other solutions rely on exploiting some side information obtained using a lateral color camera [3]. In addition, other techniques are employed to interpolate the depth maps in order to increase the resolution and fill missing data [4].

The paper presents a joint denoise-interpolation algorithm for MS Kinect sensor that aims at correcting the computed depth values and interpolate the depth map on those points where depth values are not available because of the noise conditions. The approach relies on an initial denoising performed by matching borders between range and color images. Then, a segmentation of the color image is employed to interpolate the data. Experimental results show that the quality of the processed range image improves both in terms of number of available points and quality of the warped views.

In the following, Section 2 presents the structure of the MS Kinect sensor. Section 3 describes the denoising and interpolation algorithm in detail, with Subsection 3.3 reporting the depth correction algorithm and Subsection 3.6 showing how segmentation is employed to interpolate data. Experimental results (Section 4) and conclusions (Section 5) end the paper.
2. A SHORT DESCRIPTION OF THE INFRARED STRUCTURED LIGHT SENSOR

To test the proposed approach we employed the MS Xbox Kinect device, a low-cost 3D sensor that is available on the market and exploits an IR structured light camera to estimate depth signals for the acquired scene. Despite this, the approach can be applied to any IR-based range camera. Figure 1 shows a simplified block diagram of the device. The implemented IR depth sensor consists in an IR projector, an IR CMOS camera, and a processing unit that controls them and elaborates the acquired signal. An IR pattern of dots is projected by the IR projector on the scene, and the IR CMOS camera acquires the reflected pattern, which will be distorted according to the geometry of the objects. The central processing unit estimates the distance of each point from the IR camera considering the distortions in the acquired dot pattern and elaborates the acquired signal. An IR pattern of dots is projected by the IR projector on the scene, and the IR CMOS camera acquires the reflected pattern, which will be distorted according to the geometry of the objects. The central processing unit estimates the distance of each point from the IR camera considering the distortions in the acquired dot pattern with respect to the projected one. Color information is available as well since an RGB CMOS camera permits obtaining a standard picture of the acquired scene.

This information permits building a pointcloud model of the 3D scene by mapping depth pixels into color pixels with a warping operation. The obtained 3D model presents several artifacts depending on possible calibration errors, lighting conditions and errors in depth estimation by the processing unit.

3. STRUCTURE OF THE PROPOSED ALGORITHM

The structure of the algorithm is summarized in Fig. 2 and consists in two main operating blocks: a denoising unit that corrects mismatches between the color image \( I_{in} \) and the warped depth map \( D_{in} \), and an interpolating strategy that fills holes and missing pixels in the range image.

The following subsections will describe each step in detail.

3.1. Clustering depth values

At the beginning of the depth correction strategy, the depth values \( D_{in}(x, y) \) (where \( (x, y) \) are the pixel coordinates) are clustered into a set of 20 classes \( C_k, k = 0, \ldots, 19 \), according to their distance from the IR camera using the k-means algorithm. The choice of using k-means algorithm and computing 20 classes was driven by the need of having a low complexity architecture.

Each class is characterized by its centroid and two threshold values that defines the upper and the lower bounds for depth values, that are grouped into the set of thresholds \( Th \).

3.2. Computing mismatches

At the beginning of the error correction unit, \( 3 \times 3 \) Sobel operators \( S_x \) and \( S_y \) are applied to both the luminance component \( L \) of the color image and the warped depth image \( D_{in} \). Let \( S_x \ast L \) and \( S_y \ast L \) be the convolutions of \( L \) with the horizontal and the vertical Sobel operators, respectively, and \( S_x \ast D_{in} \) and \( S_y \ast D_{in} \) be the convolution of the same operators with \( D_{in} \). Then, the edge images \( E_x \) and \( E_D \) are computed as:

\[
E_L = \text{round} \left( \frac{1}{64} \left| S_x \ast L \right| + \left| S_y \ast L \right| \right)
\]

\[
E_D = \text{round} \left( \frac{1}{8} \left| S_x \ast D_{in} \right| + \left| S_y \ast D_{in} \right| \right)
\]

where quantization steps 8 and 64 have been chosen from a set of experimental trials. Coordinates \( (x, y) \) have been omitted for the sake of conciseness. For all the pixel positions \( (x, y) \) such that \( D_{in}(x, y) \) is not valid, \( E_D(x, y) \) is set to 0.

In a second step, the mismatches between \( I_{in} \) and \( D_{in} \) are computed for each class \( C_k \) independently generating the pixel sets:

\[
C_k' = \{(x, y) \in C_k : E_D(x, y) > 0\}
\]

which comprises points in depth layer \( k \) with edge strength greater than 0. For each class \( C_k' \), the algorithm computes the displacement vector \( v^* = [v_x^*, v_y^*] \) in the search window \( W_{SR} \) such that:

\[
v^* = \arg \max_{v \in W_{SR}} \sum_{(x, y) \in C_k'} E_L(x \oplus v_x, y \oplus v_y).
\]

where the operator \( \oplus \) equals + or - depending on whether the coordinate \( x \) (or \( y \)) is higher or not with respect to the corresponding coordinate of the principal point \( R = (R_x, R_y) \). In this way, the algorithm compensate the mismatch between edges of the color component and of the depth information (see Figure 3). Despite object profiles result irregular in the depth component, the algorithm assumes that errors vary symmetrically with respect to borders and maximizing edge matching permits a correct alignment.

3.3. Correcting depth values

The correction of the depth values obtained by the IR structured light camera can be obtained differentiating the equa-
Fig. 3. Example of computation of \(v^*\) for the class \(C'_k\) (detail from the scene bearbins).

In order to interpolate depth information, empty segments (i.e., not including a sufficient number of samples from \(D'\)) need to be merged to one of the neighboring non-empty segments minimizing the MSE of the difference between the average color component of the two segments. The resulting segmentation map will be referenced as \(\mathcal{M}'\).

3.6. Interpolating depth values

After refining the segmentation map, the valid depth values within segment \(M'_k\) can be interpolated in order to fill the missing values within the same segment. This operation relies on the assumption that the depth signal is approximately smooth within each segment \(M'_k\). However, the borders of objects in the depth image are highly noisy and irregular, and therefore, some of the depth pixels could lie on a different segment \(M'_k\).

The interpolation strategy has to select the pixel values concerning the object covered by \(M'_k\) and discard the extraneous ones. For each \(M'_k\) in \(\mathcal{M}'\), the algorithm computes the variance of depth pixels and compares the resulting value with the threshold \(T_v\).

In case the variance is higher, it is possible that extraneous depth pixels are included in \(M'_k\) and they have to be removed. To this purpose, k-means algorithm is run on depth pixels of \(M'_k\) partitioning the values \(D'(x, y)\) into 3 classes. The algorithm will consider for the interpolation only those depth pixel within the most frequently chosen cluster. In case variance is lower, all the depth pixel in \(M'_k\) are used to fill holes.

After discriminating pixels to be interpolated, a polynomial regression is run on pertinent depth pixels, and the resulting coefficients are used to compute the missing pixel values. The resulting depth map will be named \(D_{out}\).

4. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed algorithm, we built a stereo system consisting of an MS Kinect sensor and a side standard webcam that acquires images with resolution 320 × 240. Different 3D scenarios have been acquired under different light conditions. For each scene we performed 10 independent acquisitions in order to obtain different realizations for the noise signal on depth maps. The performance of the algorithm has been evaluated computing the average number of valid pixels in the final depth maps and the average PSNR obtained warping the pixels of Kinect color camera on the view corresponding to the side webcam. All the experimental data are available at [7].

Figure 4(a) shows the average PSNR values of the warped views for the original depth map \(D_{in}\), the corrected depth map \(D'\), and the final depth map \(D_{out}\), while Figure 4(b) reports the percentage of valid pixels for the different depth maps. The displayed results report also the performance of the interpolation strategy applied directly to the input depth map \(D_{in}\) (in this case the final depth map is referenced as \(D_{out}^{\prime}\)). It is possible to notice that the interpolation strategy permits
5. CONCLUSION

The paper presents a joint denoising and interpolation approach for the MS Kinect sensor. The algorithm is based on an initial correction of depth values in the sequence, which then will be interpolated in order to fill holes and missing depth values. The proposed solution permits obtaining significant improvements for 3D models acquired under both controlled and uncontrolled light conditions.

6. REFERENCES


