MULTISCALE FRAMEWORK FOR ADAPTIVE AND ROBUST ENHANCEMENT OF DEPTH IN MULTI-VIEW IMAGERY

Hannes Helgason, Haopeng Li, and Markus Flierl

School of Electrical Engineering
KTH Royal Institute of Technology, Stockholm

ABSTRACT

Depth Image Based Rendering (DIBR) is a standard technique in free viewpoint television for rendering virtual camera views. For synthesis it utilizes one or several reference texture images and associated depth images, which contain information about the 3D structure of the scene. Many popular depth estimation methods infer the depth information by considering texture images in pairs. This often leads to severe inconsistencies among multiple reference depth images, resulting in poor rendering quality. We propose a method which takes as input a set of depth images and returns an enhanced depth map to be used for rendering at the virtual viewpoint. Our framework is data-driven and based on a simple geometric multiscale model of the underlying depth. Inconsistencies and errors in the inputted depth images are handled locally using tools from the field of robust statistics. Numerical comparison shows the method outperform standard MPEG DIBR software.

***Index Terms*—** DIBR, Free Viewpoint Television, Depth Consistency, Adaptive Estimation, Multiscale Modelling.

1. INTRODUCTION

The emerge of free viewpoint television and 3D video has raised interest in multi-view imagery, which allows users to access 3D scenes freely and interactively. A widely used technique for enabling free viewpoint experience is Depth Image Based Rendering (DIBR), which utilizes one or more reference texture images and their associated depth images to synthesize virtual camera views [1]. In essence, DIBR projects original pixels from reference images into 3D world coordinates according to their depth values as specified by the associated depth images. These coordinates are then projected onto the image plane of the virtual camera view. Thus, depth images play a crucial role in DIBR and more accurate depth maps can improve the quality of rendered views.

Many conventional depth image estimation methods infer the depth information from pairs of texture images instead of jointly considering all references [2]. Meanwhile, some depth improvement methods enhance the quality of individual depth images by applying smoothing-based methods on a single view [3]. These can lead to inconsistencies among multiple reference depth images. Such inconsistencies can be due to many factors, such as illumination differences among multiple texture references, pairwise matching of references instead of jointly matching multiple references, etc.

To enhance the consistency among multiple references, so-called multi-hypothesis-based methods have been proposed for improving the pixelwise consistency among references [4]. Other examples are feature-based improvement methods which utilize reliable image features to correct the reference depth images [5, 6].

Considering the geometrical properties of depth images (having smooth variations over connected areas and sharp discontinuities along boundaries of objects) we expect there to be regions in the targeted frame where parts of the depth images seem to agree, while corresponding data from the other depth images appear as outliers. This paper proposes an adaptive, multiscale, and flexible depth map estimation procedure which can deal with this situation.

The paper is organized as follows. Section 2 presents our depth enhancement framework and an evaluation of the procedure arising is given in Section 3. We finish with some discussion and future perspectives in Section 4.

2. METHODOLOGY

2.1. Problem statement and data model

Assume we have K reference depth images for a target camera view. Let I be the set of pixels in the corresponding n-by-m rectangular camera frame whose total number of pixels is denoted by $N = n \cdot m$. Here we focus on the square case $n = m$ (the methodology could be extended to cases where $n \neq m$). We take the image domain as the continuous square $[0, 1]^2$ so that the pixels in I form an array of $1/n$-by-$1/n$ squares. Furthermore, we assume that n is dyadic, $n = 2^J$, for an integer J.

Let $d(i)$ denote the true depth value at pixel $i \in I$ and let the measurement $y^{(k)}(i)$ denote the value of depth image $k = 1, \ldots, K$ at pixel $i \in I$. Our goal is to estimate the true depth given the measurements. Next we make some assumptions about the data for motivating our proposed method.

Our first step is to model the measurements by

$$y^{(k)}(i) = d(i) + Z^{(k)}(i), \qquad k = 1, \ldots, K, \quad i \in I, \tag{1}$$

where $Z^{(k)}(i)$ is measurement error. According to the discussion above, the classical assumption of assuming the errors $Z^{(k)}(i)$ to be Gaussian would not suit well for describing the inconsistencies among the reference depth images.

One can observe that typical depth images for real-world scenes are "cartoon-like," where geometry plays a bigger role than texture. One could think of a depth image as a composition of several smooth regions, each corresponding to a surface of an object in the scene; the boundaries of the regions are piecewise smooth curves, corresponding to the outlines of an object obstructing the background or another object (see Fig. 1). Hence it seems reasonable to model depth as a piecewise constant functions of the form

$$\widetilde{d}(i) = \sum_{m=1}^{M} \alpha_m g_m(i), \tag{2}$$

Email addresses: {hannes.helgason,haopeng,mflierl}@ee.kth.se

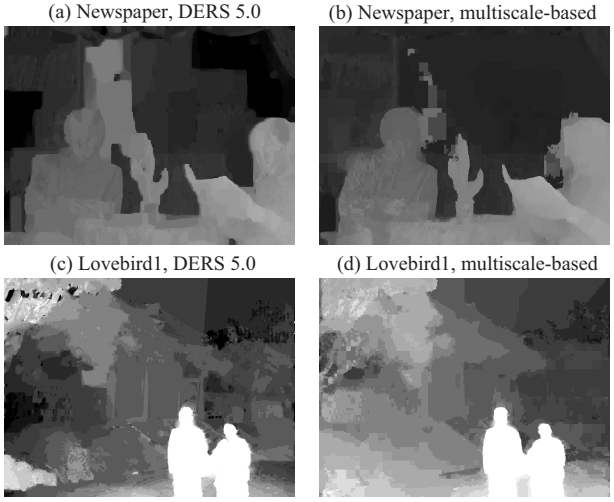

Fig. 1. Comparison of depth images at target view position.

for a real-valued sequence (α_m) and indicator functions $g_m(i) = 1_{S_m}(i)$, with $\cup_{m=1}^{M} S_m = I$ and $S_m \cap S_{m'} = \emptyset$ for $m \neq m'$.

2.2. Multiscale framework for depth enhancement

Our approach is based on models of the form (2) where the set of supports $P = \{S_m, m = 1, \ldots, M\}$ consists of *dyadic squares* forming a *recursive dyadic partition* (RDP). Here a dyadic square is a region

$$S(k_x, k_y, j) = [k_x 2^{-j}, (k_x + 1)2^{-j}] \times [k_y 2^{-j}, (k_y + 1)2^{-j}],$$

for integers k_x, k_y, j s.t. $0 \leq k_x, k_y < 2^j$ and $0 \leq j \leq J$; the pixel set I corresponds to the dyadic squares $S(k_x, k_y, J)$ for $0 \leq k_x, k_y \leq n$. We can think about the dyadic squares as nodes in a tree with $J + 1$ levels, where a dyadic square at level j is connected to the four dyadic squares at level $j + 1$ it can be decomposed into. An RDP is a partition of $[0, 1]^2$ which can be reached by the following rules: (a) $P = \{[0, 1]^2\}$ is an RDP; and (b) if $P = \{S_1, \ldots, S_M\}$ is an RDP, then a new RDP can be formed by decomposing one of the dyadic squares S_m into four dyadic squares (provided S_m is not at the bottom level $j = J$).

Our procedure can be summarized by two steps:

(i) *Local estimation:* For each dyadic square, estimate the depth based on the data associated to this square. Measure the fit of the estimate to the data and assign it as a cost to the corresponding node in the tree.

(ii) *Global estimation using model selection:* Find a balance between model's global *goodness-of-fit* to the data vs. model *complexity*.

Local estimation of depth: Our local estimate for each dyadic square is based on asserting the depth to be constant over that square. The choice of local estimation procedure depends on the nature of the error terms in the data model (1). For example, the standard assumption of taking the error terms to be i.i.d. zero-mean Gaussian would under the constant depth hypothesis lead to an estimate which is the average of all the measurements which fall in the dyadic square – the natural criterion for goodness-of-fit would be the squared ℓ_2-distance. However, as argued above, the Gaussian assumption does not seem appropriate for the data under consideration. Instead, we will view the data as a set of measurements with outliers, which brings the local estimation problem into the field of *robust statistics* [7]. Here we make a simple choice and take the popular *median* as our robust local estimate. Hence the local depth estimate $\widehat{\alpha}_m$ corresponding to the dyadic square S_m is

$$\widehat{\alpha}_m := \text{Median}\{y^{(k)}(i) : i \in S_m, k = 1, \ldots, K\} \quad (3)$$

$$= \operatorname*{argmin}_{\alpha_m} \sum_{k=1}^{K} \sum_{i \in S_m} |y^{(k)}(i) - \alpha_m|. \quad (4)$$

Thus the natural goodness-of-fit measure for the median is the ℓ_1-distance so we define the local fit for dyadic square S_m as

$$C(S_m \mid y) := \sum_{k=1}^{K} \sum_{i \in S_m} |y^{(k)}(i) - \widehat{\alpha}_m|. \quad (5)$$

Global estimation and model selection: To motivate our choice of criterion for the model selection procedure, we consider the generalized likelihood principle where the measurement errors are assumed to be i.i.d. and Laplace distributed (due to its fat tails, the Laplace distribution $f(x) = e^{-|x|}/2$ is often used to model outliers). This leads to taking

$$\min_{(\alpha_m)} \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{i \in S_m} |y^{(k)}(i) - \widetilde{d}(i)| = \sum_{m=1}^{M} C(S_m \mid y)$$

as the goodness-of-fit measure for the model (2). We choose to measure the model complexity by the number of terms M in the model (2) and define the complexity-penalized functional, for parameter $\lambda > 0$, as

$$J_\lambda(P) := \sum_{m=1}^{M} C(S_m \mid y) + \lambda M = \sum_{m=1}^{M} C_\lambda(S_m \mid y),$$

where $P = \{S_m, m = 1, \ldots, M\}$ is an RDP and $C_\lambda(S_m \mid y) := C(S_m \mid y) + \lambda$. Our proposed global estimator is

$$\widehat{d}_\lambda(i) = \sum_{S_m \in P_\lambda^*} \widehat{\alpha}_m 1_{S_m}(i); \quad P_\lambda^* := \operatorname*{argmin}_P J_\lambda(P), \quad (6)$$

where the minimization is taken over all RDPs in our tree. The choice of parameter λ depends on the balance between the richness in the structure of the underlying depth and how severe the errors in the data can be: small λ are preferred for capturing highly variable depth but large λ for fighting inconsistency (λ also scales with K and N). The examples we consider in Section 3 indicate that the procedure performs well over a range of λ – a practical procedure for choosing λ adaptively or by experience, for different types of scenes and camera setups, is a subject of future research.

2.3. Algorithms and computational complexity

In the local estimation step above, we need to take the median over all the dyadic squares in the tree. Due to the recursive structure of the tree, this can be done efficiently using a simple extension of the *merge sort* algorithm for sorting lists [8]. The computational complexity is $O(KN \log N)$, where $N = n^2$ is the total number of pixels. Now there are $O(\log N)$ levels and at each level j there are 4^j squares with $N4^{-j}$ pixels each, so once we have the medians, calculating the local costs (5) for all dyadic squares requires $O(KN \log N)$ operations.

Table 1. Settings for Test Sequences.

Sequence name	Target camera	Reference cameras	Multiscale-based depth references
Newspaper	4	2,6	2,3
(Indoor)	4	3,5	5,6
Lovebird1	6	4,8	4,5
(Outdoor)	6	5,7	7,8

To find the estimator (6), we first decorate the nodes in the tree with the penalized costs $C_\lambda(S_m \mid y)$. Then one can minimize complexity functional $J_\lambda(P)$ over all possible RDPs using a standard tree optimization algorithm of complexity $O(N \log N)$ (see [9, 10]). Hence the overall complexity for our procedure is $O(KN \log N)$. Compare this to a simple *pixelwise median* approach which estimates depth by taking medians over the warped reference depth images pixel by pixel – the complexity of such procedure is $O(KN)$.

3. NUMERICAL EXPERIMENTS

We compare the performance of our proposed method to that of the View Synthesis Reference Software 3.5 (VSRS 3.5) [11] which is used for MPEG 3DV/FTV exploration experiments. VSRS 3.5 uses a DIBR approach which synthesizes the target view by referencing left and right texture images and their associated depth images. The reference depth images are generated by MPEG 3DV/FTV Depth Estimation Reference Software 5.0 (DERS 5.0) [2]. We also consider the *pixelwise median* approach mentioned in Section 2.3 – this is essentially what the proposed estimator (6) would give for $\lambda = 0$.

The luminance PSNR (Y-PSNR) between rendered view and corresponding actual camera view is used for evaluating an objective performance for each method. We use the multi-view video test sets *Newspaper* (provided by GIST [12]) and *Lovebird1* (provided by ETRI [13]); the resolution of the videos is 1024×768 and we use 50 successive frames; the setting for the test sequences is in Table 1. Note that VSRS 3.5 only needs two references to synthesize the target view (col. 3 in Table 1) – for a fairer comparison, we use the same texture references for texture warping for all the methods. We use inverse-mapping to warp the reference textures to the target position.

For the multiscale-based method, we first warp the reference depth images listed in col. 4 of Table 1 to the target position. We divide each frame into 4×3 square blocks of size 256×256 each. The depth is estimated independently on each block using (6). This is done for parameter values λ ranging from 2 to 1000.

3.1. Improvement of depth image

Fig. 1 shows comparisons of depth images at target position between DERS 5.0 and the proposed multiscale depth enhancement approach. The geometry information appears to be better explained by the multiscale-based approach than by DERS 5.0.

3.2. Objective performance comparison

As shown in Fig. 2, our multiscale-based algorithm for depth image improvement outperforms the MPEG reference algorithm for a wide range of λ values. The average Y-PSNR of the rendered images improves by about 2dB for the outdoor sequence (*Lovebird1*) and about 1dB for the indoor sequence (*Newspaper*). The pixelwise median performs surprisingly well but the results for the *Newspaper* sequence support that it is worth going multiscale.

(a) Newspaper, VSRS 3.5 (b) Newspaper, multiscale-based

(c) Lovebird1, VSRS 3.5 (d) Lovebird1, multiscale-based

Fig. 3. Comparison of rendered images at target view position.

3.3. Subjective performance comparison

The comparison of subjective quality of rendered images are depicted in Fig. 3. The object details appear more piecewise smooth and visually better for the proposed method as it improves the multi-view consistency on a multiscale basis.

4. CONCLUSIONS AND PERSPECTIVES

We proposed a simple data-driven multiscale method for improving quality of depth maps in a robust manner and a practical multi-view video setting. Results from numerical demonstration are promising and the procedure has low computational complexity and memory requirements.

We envisage extending and evaluating the framework using more sophisticated geometric multiscale tools –such as wedgelets and platelets [14, 15]– for better modeling the underlying depth; that is, use different structure of support sets S_m for the indicator functions g_m in (2). Such geometric modeling seems very relevant for depth image coding and has been used for joint bit allocation for texture images and associated depth (see [16] and references therein). In relation to this, we point out that the estimation procedure (6) provides a depth map which is already in a condensed form – data compression is, in a way, a byproduct of the method.

For the local estimation one might want to consider other types of robust estimators than (3), such as M-estimators, trimmed means, and others [7]. The reason is that some pixel regions might not have outliers (or inconsistencies) so the median could be overly conservative and an unnecessarily high price is paid for robustness. Instead one might want to construct data-driven robust estimators which behave as closely as classical estimators –such as simple averaging– when outliers are absent or few in number. An interesting research direction could be to consider consistency testing for local groups of pixels. The motivation is that one might expect some cameras to be consistent over small regions of neighboring pixels. (Locally, this would in some sense extend ideas from [4] which work pixelwise.) One approach for weeding out outliers in depth measurements in a

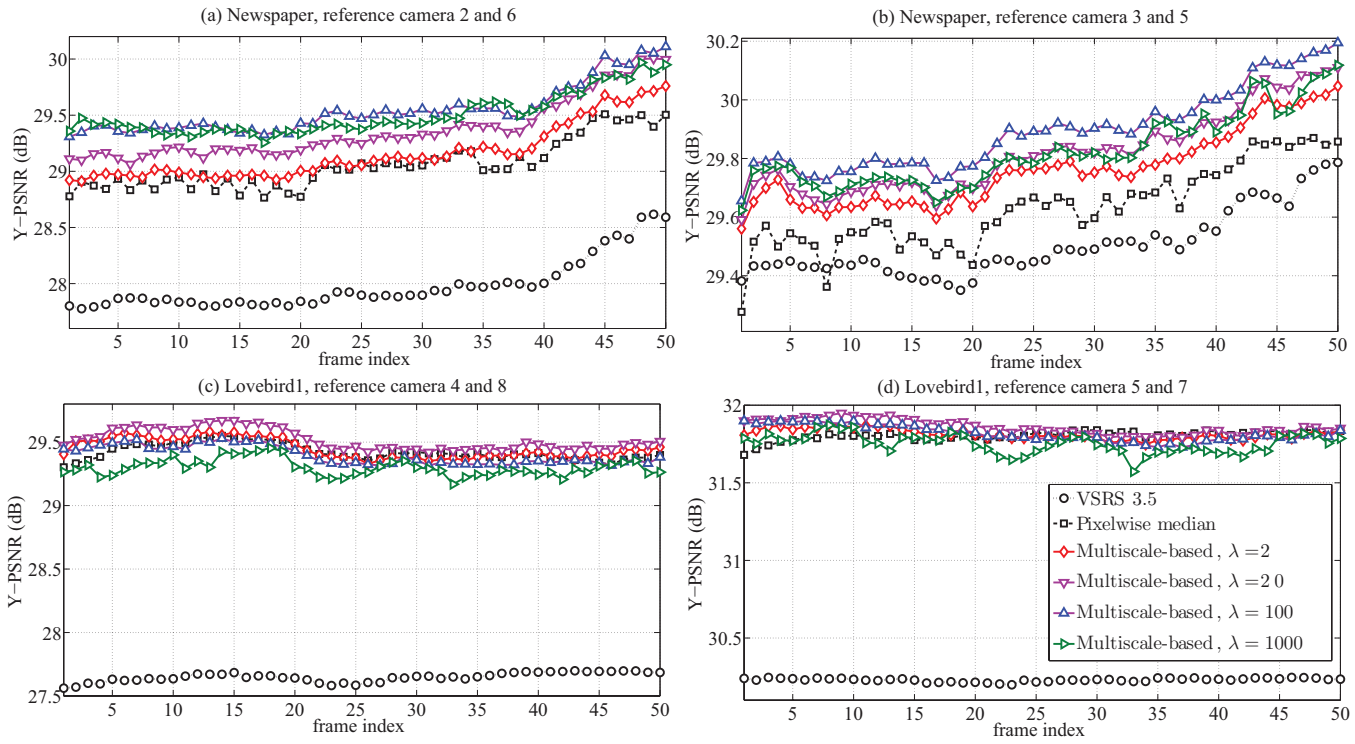

Fig. 2. Objective performance comparison between proposed and references algorithms. The labels of the curves in subplots (a), (b) and (c) are the same as those of (d). (Note the different range and scale for the y-axes.)

dyadic square S at level j could be to look for a set $A \subset S$ of pixels, with a targeted size $|A| = L = L(j)$ (say, half of the total number of measurements in S), which minimizes the range

$$\max_{i,i' \in A,\ k,k'=1,\ldots,K} |y^{(k)}(i) - y^{(k')}(i')|.$$

Finally, we would like to mention that robust local estimation procedures could provide information about local consistency among the reference depth maps – this could then be used in DIBR for choosing which reference texture images to use for local image regions.

5. REFERENCES

[1] Y. Mori, N. Fukushima, T. Fujii, and M. Tanimoto, "View generation with 3D warping using depth information for FTV," *3DTV Conference*, pp. 229–232, May 2008.

[2] M. Tanimoto, T. Fujii, and M. Panahpour, "Depth estimation reference software DERS 5.0," Rep. M16923, Oct. 2009.

[3] W. J. Tam, G. Alain, L. Zhang, T. Martin, R. Renaud, and D. Wang, "Smoothing depth maps for improved stereoscopic image quality," in *Three-Dimensional TV, Video and Display III*, Oct. 2004, pp. 162–172.

[4] P. Rana and M. Flierl, "Depth consistency testing for improved view interpolation," in *Proc. of the IEEE Workshop on Multimedia Signal Processing*, Oct. 2010, pp. 384–389.

[5] S. Kumar, M. Kumar, B. Raman, N. Sukavanam, and R. Bhargava, "Depth recovery of complex surfaces from texture-less pair of stereo images," *Electronic Letters on Computer Vision and Image Analysis*, vol. 8, no. 1, pp. 44–56, 2009.

[6] H. Li and M. Flierl, "SIFT-based improvement of depth imagery," in *Proc. of the IEEE International Conference on Multimedia & Expo*, Barcelona, Spain, July 2011, pp. 1–6.

[7] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, Wiley, New York, 2nd edition, 2009.

[8] D. E. Knuth, *Art of Computer Programming, Vol. 3: Sorting and Searching*, Addison-Wesley, 2nd edition, 1998.

[9] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. Info. Theory*, vol. 38, pp. 713–718, 1992.

[10] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press Inc., San Diego, 1999.

[11] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis," Rep. M15377, Apr. 2008.

[12] Y. Ho, E. Lee, and C. Lee, "Multiview video test sequence and camera parameters," Rep. M15419, Apr. 2008.

[13] G. Um, G. Bang, N. Hur, J. Kim, and Y. Ho, "3D video test material of outdoor scene," Rep. M15371, Apr. 2008.

[14] D. Donoho, "Wedgelets: Nearly minimax estimation of edges," *Ann. Statist.*, vol. 27, pp. 859–897, 1999.

[15] R. M. Willett and R. D. Nowak, "Platelets: a multiscale approach for recovering edges and surfaces in photon–limited medical imaging," *IEEE Transactions on Medical Imaging*, vol. 22, no. 3, pp. 332–350, 2003.

[16] Y. Morvan, *Acquisition, Compression and Rendering of Depth and Texture for Multi–View Video*, Ph.D. thesis, Technische Universiteit Eindhoven, 2009.