# Depth map upsampling and refinement for FTV systems

Adrian Dziembowski, Adam Grzelka, Dawid Mieloch, Olgierd Stankiewicz, Marek Domański
Chair of Multimedia Telecommunications and Microelectronics,
Poznan University of Technology, Poland
{adziembowski, agrzelka, dmieloch, ostank}@ multimedia.edu.pl

*Abstract* — **In free-viewpoint television, high-quality depth maps are substantial for a virtual view synthesis for free navigation purposes. In this paper, we propose a new method of depth map quality improvement and resolution increase. Our method is intended for low resolution depth maps acquired through estimation using multiview video. The proposed depth map quality improvement is based on segmentation of an acquired high-resolution image. In the estimated depth map, searching for outliers is performed in the neighborhood created from similar segments in the acquired view. Experimental results show high effectiveness of the presented method of depth maps refinement, especially for object edges.**

***Keywords — depth map upsampling; depth map refinement; segmentation; free-viewpoint television, virtual navigation***

## I. Introduction

Recently, the demand for development of new video-based solutions related to Augmented Reality and Virtual Reality is rapidly growing. For the video technology, this implies a need to develop the efficient tools for virtual navigation within natural 3D scenes. Such functionality will be provided by Free-viewpoint Television (FTV) [1]. A viewer of such interactive video can move his/her virtual viewpoint at any time. The viewpoints available for an FTV viewer are not limited to the locations of the real cameras located around a scene, but any virtual view from any viewpoint around a scene can be created using the virtual view synthesis. Potential applications of future FTV systems include sports and performance broadcastings as well as interactive courses and teaching materials.

The most challenging step towards the virtual view synthesis is depth estimation [2]. For a view, the depth map can be obtained using the depth sensors (e. g. Microsoft Kinect, time-of-flight cameras). Unfortunately, the usage of such active sensors in multiview systems is problematic because of possible interferences between multiple sensors and their limited usability for outdoor scenes. Software depth estimation is free of aforementioned flaws, nevertheless it is very time consuming. One of the possibilities is depth estimation from the views with reduced resolution. Unfortunately, the quality of the synthesized views for the low-resolution depth maps is moderate, especially near the edges. Therefore, we propose a fast method for depth maps upsampling and refinement in order to achieve high quality of synthesized virtual views

## II. Related work

The concept of depth map upsampling or refinement has been already studied because of the difficulty and the complexity of estimation the depth maps with high resolution. The largest group of existing algorithms focuses mainly on depth map upsampling [3-5].

The methods [4,5] are based mainly on multi-lateral and multi-dimensional filtering which gives a possibility of very efficient real-time implementations. However, this approach is not very efficient for algorithmically estimated depth maps. The filtering highly reduces a noise on noisy fragments of a depth map but does not recover a depth of objects incorrectly assigned to a background of the scene. Such methods are better suited for the use with depth sensors.

Other algorithms [6,7] focus on a filling of the holes in the depth maps. Such holes result from occlusions present in the acquired scene. Also aforementioned method [3] uses segmentation to inpaint holes in depth maps. However, the depth map completion is especially needed when depth sensors are used. In the spatially consistent multi-view depth map estimation [8] the number of occluded regions is highly reduced, therefore, the process of depth map holes filling is not required

An interesting method [9] is based on matching the segments in stereopair to reconstruct high-quality and upscaled depth maps. A similarity and a size of corresponding segments in the left and the right view of stereopair is very high because of the small distance between cameras and their parallel optical axes. Unfortunately, this is not true for sparse arrangement of arbitrally located cameras.

## III. Depth map estimation algorithms

Before the proposed method of upsampling and refinement of depth maps will be described it is necessary to introduce the basics and techniques of depth map estimation.

In order to calculate the depth of a point, it has to be simultaneously visible by at least two cameras [2]. In the simplest methods, called local estimation methods, the process of the depth estimation is performed independently for all points of an acquired image. For each point a search for most similar corresponding point in another view is performed. The

The presented work has been funded by the Polish Ministry of Science and Higher Education for the status activity consisting of research and development and associated tasks supporting development of young scientists and doctoral students.

point in another image with the highest similarity to actually processed point is chosen as the best and used for depth calculation. Methods in this group usually differ in used points similarity metric. Independent calculations for each point make it possible to estimate the depth in real time. Nevertheless, the quality of resulting depth maps is too low for virtual view synthesis purposes because of no inter- and intra-view consistency of depth maps. Estimated depth maps are highly noised.

The second group of depth estimation techniques is based on global optimization algorithms (e.g. Graph Cut [10] or Belief Propagation [11]). In performed optimization an objective function is usually expressed as a sum of previously described similarity metric and smoothness between neighboring points, summarized for all points of all acquired views [10]. This approach results in much higher quality of estimated depth maps than in local estimation methods, preservation of consistency between cameras and a smooth depth on surfaces of objects [2]. Naturally, the optimization process is very time consuming, especially when performed for a great number of high resolution cameras. Estimation for decimated images reduces complexity of a problem but estimated depth maps have to be additionally upsampled and refined in post-processing.

## IV. Proposed method

### A. Assumptions for the technique

First of all, the method of depth map upsampling and refinement has to increase a depth map resolution to the resolution of acquired views. Moreover, the designed method has to guarantee the shifting of depth map edges to correct positions. Depth map edge displacement errors, which are the effect of low resolution estimation or depth oversmoothing, highly reduce a subjective quality of the virtual view synthesis. Furthermore, the smooth depth of planes, which are usually already present in low resolution depth maps, should be preserved in upsampled image. On the other hand, if an area of a depth map is noised though should be smooth, the incorrect depths should be corrected.

### B. The main idea

In the first step, an acquired view corresponding to the actually processed depth map is segmented using superpixel segmentation [12]. For each segment a neighborhood from adjacent similar segments (in terms of their color) is created. In the neighborhood of the segment the depth median is calculated. Subsequently, a depth of each point is compared to the median depth of the neighborhood of corresponding segment. If a difference between depth of point and median depth is too high (threshold is based on the spatial resolution of depth map), then for processed point a depth is changed to median depth of the neighborhood. The last step of the depth processing is bilateral filtering which ensures removal of remaining depth map errors.

### C. Segmentation

Used method of segmentation should most of all preserve edges of image. Segments size has to be managable in order to achieve possibility of simultaneous correction of erroneously estimated fragments of depth map and preservation of smooth gradients on surfaces. The method has to additionally guarantee full automation of the process and reasonable time of computation. As [12] concludes, the superpixel segmentation method SLIC ensures fulfillment of aforementioned requirements.

## V. Application of the proposed method

The described method of the depth map upsampling and refinement was proposed during a development of FTV system with sparse arrangement of cameras [8]. As the first results show [13], if the cameras are distributed as a set of camera pairs, the quality of virtual view can be substantially increased. Non-uniform arrangement of cameras allows joins advantages of small base (less occlusions) and large base (high spatial resolution of depth map). However, errors resulting from the estimation of a depth map in low resolution still remain. Hence, we propose the new method of depth map upsampling and refinement.

## VI. Design of the experiment

In order to evaluate the performance of the proposed method of depth maps upsampling and refinement an experiment was performed. Test sequences set consisted of 5 sequences: 3 natural sequences in 1920×1080 resolution (Poznan_Fencing2, Poznan_Blocks2, Poznan_Service2 [14]) and 2 synthetic ones in 1280×768 resolution (Big Buck Bunny Flowers and Big Buck Bunny Butterfly [15]). For each of test sequences 4 views arranged in two pairs on an arc of 15 degrees between a pair were used. First of all, for chosen views the depth maps were estimated in resolution four times lower than the original one, with multi-view depth estimation algorithm [8]. In the next step, estimated depth maps were upsampled and refined with the proposed method. The last step was bilateral filtering of maps.

Due to the absence of ground truth depth maps for natural sequences, the quality of depth maps was estimated indirectly through the virtual view synthesis. The synthesis of virtual view placed in real view positions was done using two neighboring views and their corresponding depth maps. The resulting virtual view was compared with real view using PSNR metric (for luminance). For each sequence two syntheses were done: from view 0 and 2 to position of view 1 and from view 1 and 3 to position of view 2. The final quality was the result of average of quality of these two syntheses.

## VII. Experimental results

Fig. 1 presents fragments of originally estimated depth maps and depth maps after each of 2 steps of the proposed method: upsampling with refinement and bilateral filtering. Depth maps are presented with a corresponding fragment of view. Presented fragments of depth maps show that after the process of upsampling and refinement the edges of depth maps are better fitted to their correct positions, namely to positions of edges present in the acquired scenes. The depth of planar surfaces are still smooth, outlier points are removed. As it was shown in Fig. 2, better quality of depth maps in neighborhood of edges highly reduces synthesis artifacts unpleasant for a viewer of free navigation video.

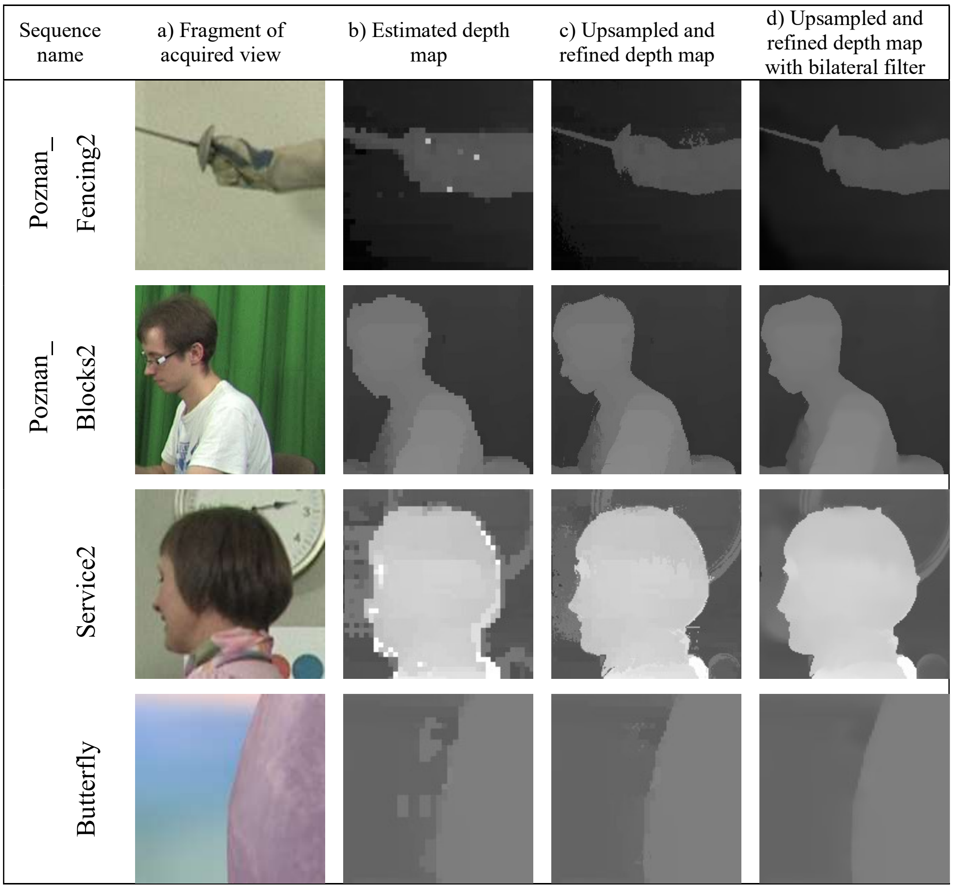
Sequence name
a) Fragment of acquired view
b) Estimated depth map
c) Upsampled and refined depth map
d) Upsampled and refined depth map with bilateral filter
Poznan_ Fencing2
Poznan_ Blocks2
Service2
Butterfly

Fig. 1. Comparison of acquired views fragments (a) to estimated depth maps (b), depth maps after upsampling and refinement (c) and depth maps after bilateral filtering (d).

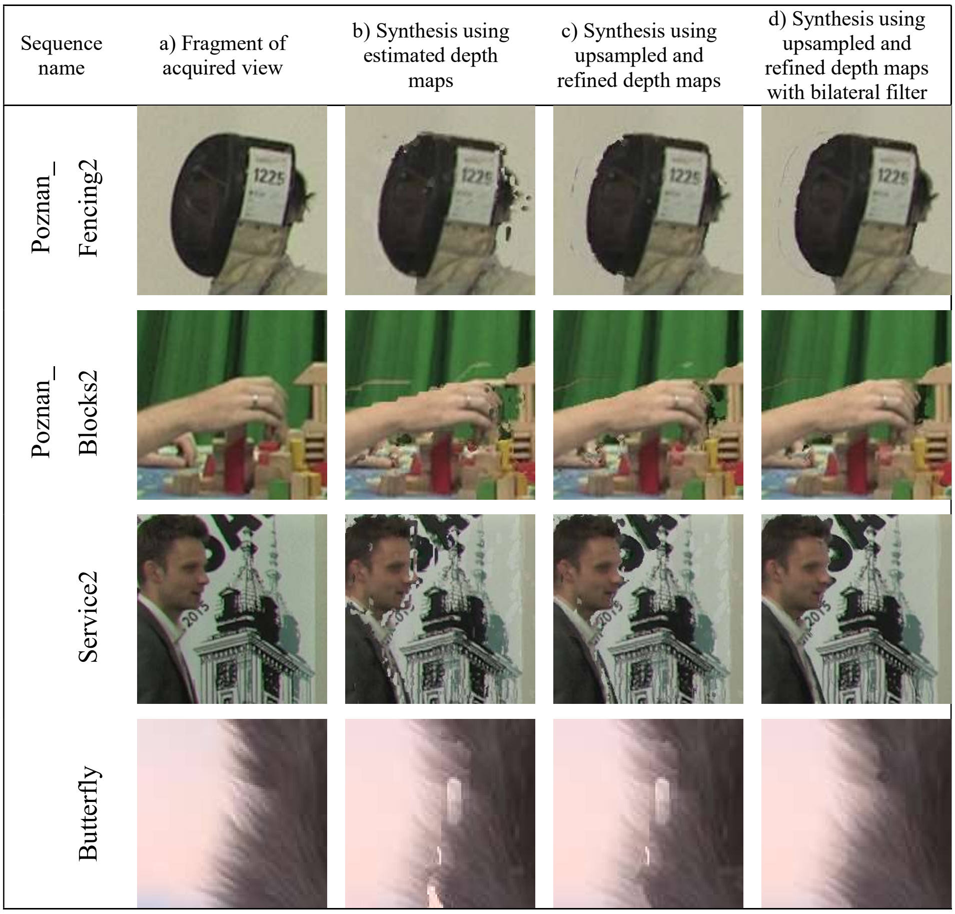
Sequence name
a) Fragment of acquired view
b) Synthesis using estimated depth maps
c) Synthesis using upsampled and refined depth maps
d) Synthesis using upsampled and refined depth maps with bilateral filter
Poznan_ Fencing2
Poznan_ Blocks2
Service2
Butterfly

Fig. 2. Comparison of acquired views fragments to fragments synthesized using: (a) estimated depth maps (b), depth maps after upsampling and refinement (c) and after bilateral filtering (d).

Table I and Figure 3 present results of objective quality assessment. For all test sequences the quality of virtual views was higher after depth map upsampling and refinement. The difference in PSNR is not high because the proposed method improves the quality of depth map edges, which usually constitute a low percentage of the depth map area.

TABLE I. VIRTUAL VIEW SYNTHESIS QUALITY USING A) ESTIMATED DEPTH MAPS, B) DEPTH MAPS WITH UPSAMPLING AND REFINING C) DEPTH MAPS WITH UPSAMPLING, REFINING AND BILATERAL FILTERING

| **Test sequence** | **Virtual view synthesis quality [dB]** | | | |
|---|---|---|---|---|
| | ***a*** | ***b*** | ***c*** | ***c-a*** |
| Poznan_Blocks2 | 27.49 | 27.60 | 27.67 | **0.18** |
| Poznan_Fencing2 | 26.95 | 27.61 | 27.75 | **0.80** |
| Poznan_Service2 | 25.16 | 25.45 | 25.77 | **0.61** |
| BBB Flowers | 18.84 | 19.23 | 19.41 | **0.57** |
| BBB Butterfly | 27.30 | 27.58 | 28.13 | **0.83** |

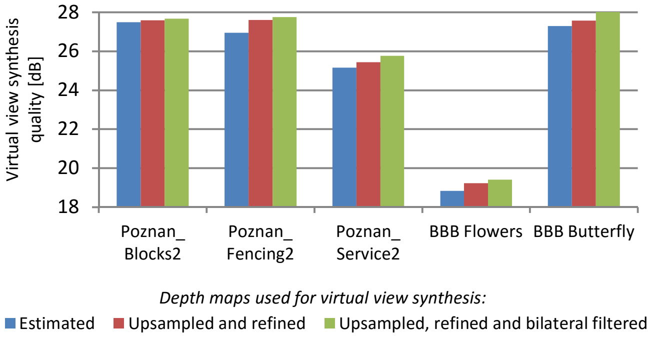

Fig. 3. View synthesis quality comparison.

## VIII. CONCLUSIONS

In this paper, we present a new method for upsampling and refinement of low-resolution depth maps. As the experiments show, upsampling and refinement increase the objective quality of virtual views for both synthetic and natural test sequences. Furthermore, better representation of depth of objects present in the acquired scene, especially in the neighborhood of the edges, significantly reduces virtual the errors of the view synthesis.

The proposed method does not impose any requirements on camera positioning, thus can be used with any camera arrangement, e.g. with stereo-pair or with Free-Viewpoint Television systems with arbitrary locations of cameras.

The method of upsampling and refinement allows restoring of correct depth even for depth maps with resolution 4 times lower in each direction than the original. The possibility of depth maps estimation for decimated resolution significantly reduces required time of computation, which is one of the last obstacles in further development of practical FTV systems.

## REFERENCES

[1] M. Tanimoto, M. Tehrani, T. Fujii, T. Yendo, “FTV for 3-D spatial communication”, Proc. IEEE, Vol. 100, pp. 905-917, April 2012.

[2] O. Stankiewicz, “Stereoscopic depth map estimation and coding techniques for multiview video systems”, PhD Dissertation at Poznan University of Technology, Faculty of Electronics and Telecommunications, 2014.

[3] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, I. Kweon, “High quality depth map upsampling and completion for rgb-d cameras”, IEEE Transactions on Image Processing, 23(12), pp. 5559–5572, 2014.

[4] J. Dolson, J. Baek, C. Plagemann and S. Thrun, "Upsampling range data in dynamic environments", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1141-1148, 2010.

[5] D. Chan, H. Buisman, C. Theobalt, S. Thrun, “A noise-aware filter for real-time depth upsampling”, Proceedings of ECCV workshop on Multi-camera and multi-modal Sensor Fusion Algorithms and Applications, 2008.

[6] P. Buyssens, M. Daisy, D. Tschumperle, O. L´ezoray, “Superpixel-based depth map inpainting for RGB-D view synthesis”, IEEE International Conference on Image Processing (ICIP), pp. 4332-4336, 2015.

[7] M. Ciotta, D. Andourtsos, “Depth guided image completion for structure and texture synthesis”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1199-1203, 2016.

[8] M. Domański, A. Dziembowski, D. Mieloch, A. Łuczak, O. Stankiewicz, K. Wegner, „A practical approach to acquisition and processing of free viewpoint video,” Picture Coding Symposium PCS, pp. 10-14, 2015.

[9] J. Cai, C. Jung, “Image-guided depth propagation using superpixel matching and adaptive autoregressive model”, Visual Communications and Image Processing (VCIP), pp. 1-4, 2015.

[10] V. Kolmogorov, R. Zabih, “What energy functions can be minimized via graph cuts?”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(2), pp. 147-159, 2004.

[11] J. Sun, N. N. Zheng, H. Y. Shum, “Stereo matching using belief propagation”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(7), pp. 787-800, 2003.

[12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, “SLIC superpixels compared to state-of-the-art superpixel methods”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11), pp. 2274-2282, 2012.

[13] M. Domański, M. Bartkowiak, A. Dziembowski, T. Grajek, A. Grzelka, A. Łuczak, D. Mieloch, J. Samelak, O. Stankiewicz, J. Stankowski, K. Wegner, “New results in free-viewpoint television systems for horizontal virtual navigation”, IEEE International Conference on Multimedia and Expo (ICME), to be published, 2016.

[14] M. Domański, A. Dziembowski, A. Grzelka, D. Mieloch, O. Stankiewicz, K. Wegner, “Multiview test video sequences for free navigation exploration obtained using pairs of cameras”, ISO/IEC JTC1/SC29/WG11, Doc. M38247, Geneva, 2016

[15] P. Kovacs, “[FTV AHG] Big Buck Bunny light-field test sequences”. ISO/IEC JTC1/SC29/WG11, MPEG Doc. M35721, Geneva, 2015.