Multiresolution example-based depth image restoration

Ljubomir Jovanov, Aleksandra Pižurica and Wilfried Philips

Telecommunications and Information Processing Department, Ghent University,
Sint Pietersnieuwstraat 41, 9000 Ghent, Belgium;

ABSTRACT

In this paper we present a new method for superresolution of depth video sequences using high resolution color video. Here we assume that the depth sequence does not contain outlier points which can be present in the depth images. Our method is based on multiresolution decomposition, and uses multiple frames to search for a most similar depth segments to improve the resolution of the current frame. First step is the wavelet decomposition of both color and depth images. Scaling images of the depth wavelet decomposition, are superresolved using previous and future frames of the depth video sequence, due to their different nature. On the other side wavelet band are improved using both previous frames of the wavelet bands and wavelet bands of color images since similar edges might appear in both images. Our method shows significant improvements over some recent depth images interpolation methods.

Keywords: depth images, wavelets, noise estimation, denoising

1. INTRODUCTION

Reliable and clean features are neccessary conditions for complex tasks such as object recognition, autonomous navigation of robots, industrial inspection and biometric authentication. Algorithms which aim to solve these problems often rely on luminance, color and motion information in order to get an interpretation of the scene. The above-mentioned features are often not sufficient for a valid interpretation of the scene due to the occlusions and the lack of information needed for a unique interpretation.

Scene interpretation can be significantly improved by introducing range data into the feature set. Depth information makes the task of the scene interpretation more feasible and robust. Various range measuring techniques exist which are based on usage of multiple cameras. These include triangulation systems such as stereo vision (or structured light), depth-from-focus, depth-from-shape and depth-from-motion. Most recent depth sensors are based on the measuring of time of flight of the light beam. This type of depth sensors offers better accuracy, higher frame rate and lower computational requirements in order to reconstruct depth image. However the main advantage of the time of flight sensors, is that they physicaly measure the distance between the camera and the object, while other techniques use image analysis, which is often prone to errors, to obtain depth images. The main disadvantage of time of flight depth sensors is their limited spatial resolution which is much below the resolution of modern video cameras. This problem was first adressed in the papers of Diebel,[1] Riemens[2] and Schuon.[3] In this paper we address this disadvantage and show that the superresolution can significantly improve the quality of the depth images.

Further author information: (Send correspondence to Ljubomir Jovanov)
Ljubomir Jovanov: E-mail: ljj@telin.ugent.be, Telephone: +32 9 264 34 12, A. Pižurica is a postdoctoral researcher of FWO Flanders, Belgium

Applications of Digital Image Processing XXXII, edited by Andrew G. Tescher, Proc. of SPIE
Vol. 7443, 74431C · © 2009 SPIE · CCC code: 0277-786X/09/$18 · doi: 10.1117/12.826226

Time of flight cameras measure depth by emitting a modulated beam of infrared light and measure the phase difference between reference light beam and the beam reflected from the scene. Since the depth is sensed physically, the whole measurement proccess is independent of of the texture in the scene (which is not the case for stereo systems that rely on disparity estimation). Typical resolutions of time-of-flight cameras available on the market today are at most 320x240 pixels. In addition, depth images obtained using time-of-flight sensors are often contaminated by noise and other errors, such as outliers.

Depth images usually have resolution which is not better than 176x144 pixels, if they are produced by time of flight camera, or 90x72 if they are produced using real-time block-matching disparity estimation algorithm eg. the real time estimator of De Haan.[4] For 3D TV transmision that employs on depth image based rendering, resolutions of color video sequence and corresponding depth have to be the same in order to have satisfactory quality of the rendered stereo images.

First methods that used high resolution color cameras were presented in the papers of Diebel[1] and Yang.[5] There authors assume that depth and luminance discontinuities are aligned, which allowed them to enhance edges using high resolution luminace or color images, while using smoothing in other regions. Related recent publications of Riemens[2] and Gangwal[6] use cross-bilateral filter for improvement of depth images using high resolution color image. In the recent publication Schuon et. al[3] adopt the method of Farsiu[7] based on L_1 norm minimization and robust regularization based on a bilateral prior and apply it directly on a sequence of depth images without introducing information from high resolution color sequence.

Initial approaches used depth upsampling which was regularized using edge consistency term with respect to the color image. In the approach of Diebel[1] authors used Markov random field for regularization of the superresolved images. Recently in the papers of Yang[5] and kopf[8] bilateral filtering was used on cost volume and color image respectively. These methods can recover well high frequency components of depth maps. However, since the *origin* of the edge in color sequence is not taken into account, whether an edge originates from the real discontinuity in depth, like object boundary or from texture alone. Consequently, disturbing artefacts tend to arise on textured flat regions of the scene. The *improved* approach of Lindner[9] improves the above mentioned problem by taking into account noise and the edges.

In this paper we present a method that uses multiresolution representation of both depth and color sequence in order to better asses the existence of edges and locate the most similar patches in the previous and following frames. Our method builds up on non-local superresolution methods like.[10] The main contribution of this paper is combining information from both domains to improve the resolution of the depth sequence.

In Section 2, we give an overview of related work on superresolution of depth and color video sequences, Section 3 gives a short introduction about multiresolution methods, Section 4 describes the proposed method. Experimental results are given in Section 5 and conclusions in Section 6.

2. RELATED WORK

Superresolution is one of the essential problems in image processing and therefore it has been thoroughly studied in the last few decades. Only recently these techniques were applied for depth images superresolution. Common principle for all the methods that were developed is that low resolution images are observed as degraded (i.e. subsampled and blurred) samples of high-resolution scene as shown in Fig.1, and combined using correspondences and weights to form high resolution image.

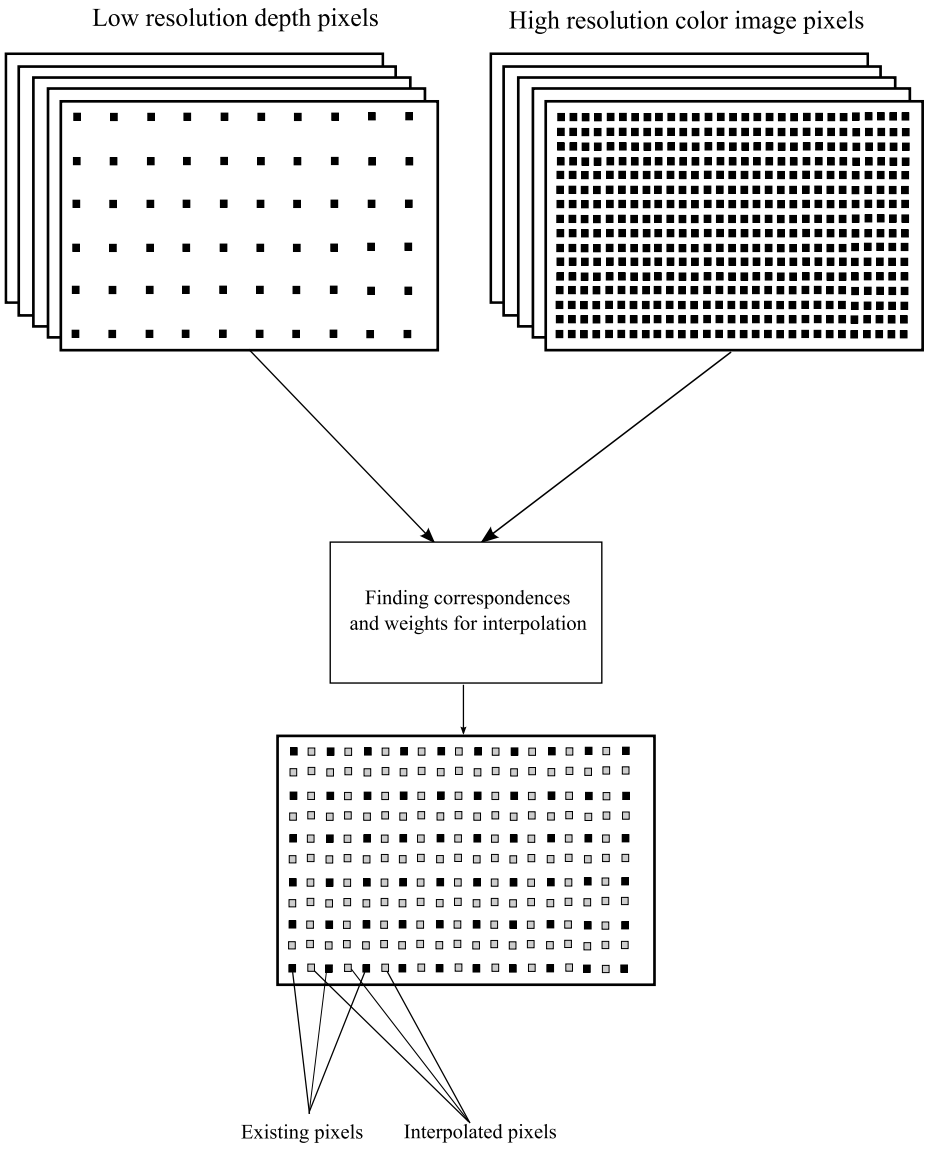

Figure 1. a.) Superresolution principle

Most of the related methods uses high resolution color images to improve the resolution of the depth image taken from the same location.

Bilateral filters were introduced by Thomasi and Manduchi.[11] Despite their simplicity, these filters offer good results in denoising and superresolution in subjective and to less extent objective measures. Basic

bilateral filter definition is

$$I'_p = \frac{\sum_{q \in S} w_{p,q} I_q}{\sum_{q \in S} w_{p,q}}, \tag{1}$$

where I_q and I'_p denote the pixel intensities at positions q and p in the input image I and the proceessed image I' respectively. In this basic configuration of bilateral filter, output image is a weighted average of pixels, where the weights $w_{p,q} = s(||p-q||)r(I_p - I_q)$ are derived from the color image pixels, and s and r are Gaussian functions with variances σ_s and σ_r. We can see from the Equation (1) that the value of the resulting pixel depends on the functions of difference in pixel values s and distance of the original and neighboring pixel r. Usually function s is a 2D convolution filter and r is the function that decreases with larger intensity differences to preserve the edges.

To apply bilateral filter for enhancement of digital photographs taken with and without flash, Eisemann[12] and Petschnigg[13] presented independently the *cross* or *joint* bilateral filter, which use the properties of one image to smooth an image with a similar content. Cross bilateral filter was first used for depth images enhancement by Kopf and Cohen.[8] The interpolated depth pixel value using cross bilateral filter is

$$D^h_p = \frac{\sum_{q \in S} w_{p,q} D^l_q}{\sum_{q \in S} w_{p,q}}, \tag{2}$$

High resolution depth map D^h is obtained by filtering the low resolution depth map D^l where the filter aperture is S guided by a high resolution texture image.

Many superresolution methods exist for color images, like[7] most of which are based on per-pixel motion estimation and regularization. Precise and consistent motion estimation is essential for these algoriths. Recently, motion-estimation free algorithms for noise reduction and superresolution[10] appeared with promising results.

The method of Schuon et al.[3] authors uses slightly displaced depth frames to obtain high resolution depth image, wwithout making use of the corresponding high resolution image. Superresolved depth images were calculated from 15 images, slightly translated orthogonally to the viewing direction. This method is derived from the approach of Farsiu et al.[7] developed for 2D digital photographs. Authors make assumption that the process of forming depth images is analogous to the image formation process of normal photo camera. Formation process of the depth image Y_k can be described by the following equation:

$$\mathbf{D^h_k} = D_k H_k F_k \mathbf{D^l_k} + \mathbf{V_k} \tag{3}$$

where $\mathbf{D^l_k}$ is the one of the low resolution depth images to be interpolated, F_k is a translation operator that represents the correspondences between the superresololved image and the current low resolution image, D_k is a decimation operator that models the downsampling from the high resolution image to the size of the low resolution image, H_k is a blur operator that models the characteristics of the lens and V_k models additive noise of the sensor. The high resolution image is estimated as follows

$$\hat{\mathbf{D}}^{\mathbf{h}} = \underset{X}{\operatorname{argmin}} [\sum_{k=1}^{N} ||D_k H_k F_k \mathbf{D^l_k}|| + \lambda \Upsilon(D^l_k)] \tag{4}$$

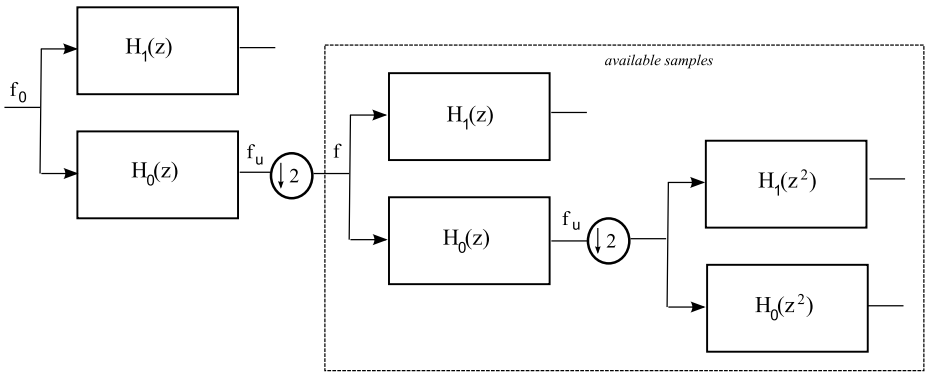

Figure 2. a.) Orthogonal wavelet transform

where $\Upsilon(D_k^l)$ is a regularization term based on bilateral prior, with corresponding weighting factor λ,

Motivated by succesful application of multiresolution techniques on image restoration problems in the papers of Pižurica[14] and Chang[15] and succesful combining of multiresolution and methods that use bilateral regularization of Zhang[16] we propose a method that combines these two concepts.

3. MULTIRESOLUTION ANALYSIS

One of the main challenges for superresolution algorithms is preservation of the edges of the objects. Unfortunately, many of the traditional superresolution and interpolation algorithms for images assume some smoothness constraints on the image, which yield blurred results. Quality of the resulting image can be significantly improved by using more sophisticated image models, that take into account the salient feature of the images, like edge-directed interpolation. One of the most important features of a signal are the points of sharp variations, or singularities. These points usually correspond to object edges or region boundaries. Edges can be efficiently detected using a multiscale analysis. For example it is well known that the Canny detector[17] is equivalent to finding local maxima of wavelet coefficients. We base part of our algorithm on these principles, since multiscale edge characterization offers convinient analysis of edges and good model for interpolation. Edges are most often extracted by examining the first derivatives of the signal or its smoothed version Local extremas of the first derivatives correspond to a regions of large variations in the image, while small magnitudes correspond to slow transitions in image. If $\theta(x)$ is smoothing function which satisfies the condition that it converges to zero when x converges to ∞ and whose integral is 1. If we denote the derivative of $\theta(x)$ as $\psi(x)$, function $\psi(x)$ can be considered as wavelet. Dilated version of a wavelet function is defined as $\psi_s(x) = \frac{1}{s}\psi(\frac{x}{s})$ where s is the scale. Wavelet transform of f at scale s and location x is given as a convolution $W_s f(x) = f(x) * (s\frac{d\theta_s(x)}{dx}) = s\frac{d}{dx}(f * \theta_s)(x)$. Taking the above definitions into account the wavelet transform of the signal at scale s is equivalent to taking the first derivative of the smoothed signal. In practice, wavelet transform is calculated by successive filtering of signal with vertically and horizontally oriented low and high pass filters as shown in Figure 2

In our method we use the sparsest level of orthogonal wavelet transform for depth sequence, since it is being interpolated, and all the other scales for color sequence that is used as reference for interpolation

Figure 3. a.) Scaling coefficients of color image b.) Scaling coefficients of luminance images c.) Horizontal wavelet band of the luminance image on a second scale d.) Horizontal wavelet band of the depth image on a second scale e.) Vertical wavelet band of the luminance image on a second scale f.) Vertical wavelet band of the depth image on a second scale

of depth sequence. Scaling coefficients of depth and color frames are quite different in nature, since they represent two different phenomenas, so we can not use higher resolution scaling images for improving lower resolution depth scaling coefficients. On the other hand, wavelet coefficient of color and depth sequence appear to be quite similar, as shown in Figure 3. The only exceptions are textured parts, which contain coefficients with high values, althought there is no real edges of objects. If taken without analysis, whether

they represent real borders of the objects, these high values can produce disturbing artefacts (i.e. non-existing edges) on flat depth regions. Otherwise, the use of wavelet coefficients significantly improves the quality of edges in depth images.

4. THE PROPOSED METHOD

In this Section we describe our proposed algorithm for superresolution of depth images. In order to make objective comparison of the algorithm performance possible, we use high resolution depth maps with corresponding luminance images, that are subsampled to obtain low resolution versions. Results of superresolution obtained using our algorithm are then compared to reference high resolution depth images. First step is wavelet decomposition of both depth and luminance information. We keep all three scales and all orientations of the wavelet decomposition of the luminance, while for depth only coarsest scale wavelet coefficients is kept, to simulate the decimation and blurring process of generating low resolution data. The main motivation to use multiresolution decomposition was found in recent succesful papers on image interpolation of Chang,[15] Velisavljevič[18] and Li,[19] combination of non-local techniques and multiresolution from Zhang[16] and succesful application of wavelet techniques on passive error concealment in paper of Rombaut.[20]

The algorithm procedes iteratively, i.e., in each iteration resolution is increased two times by calculating missing pixels. Pixels being superresolved are calculated differently for low frequency coefficients and for high frequency wavelet coefficients. The main reason is that after wavelet transform, low frequency wavelet band pixels are much more correlated and superresolution can be carried out in a simpler way, without significant artefacts. To supperresolve low frequency component of the depth images, we form spatio-temporal neigbourhood of 7 depth and luminance frames. In order to reduce computational complexity required for finding most similar neighbourhoods to the current one, we form neighbourhoods of patches of both sequences, while the size of each patch is 5x5. This is done since the memory requirements for storing patches grow quadraticaly with the increase of neighborhood. Prior to patch extraction, low pass band of all 7 depth frames are interpolated using Lancszos interpolation kernel (see the book of Turkowski[21]). Kd-tree (short for k-dimensional tree) is a space-partitioning data structure for organizing points in a k-dimensional space, that was first introduced in a paper of Lee.[22] They enable the efficient searches involving a multidimensional search key (e.g. range searches and nearest neighbor searches). Then patches of the same region from all 7 frames are formed, and k-d tree hierarchy of 50-dimensional points is formed. Then for each missing point in high resolution low frequncy band, we search for the 2 points with the smallest Euclidean distance to the current one that is being interpolated. The pixels in the low pass band (i.e. the scalling coefficients) are interpolated as follows

$$\hat{s}_d^h = \frac{\sum_{t\in[1,\ldots,T]} \sum_{(i,j)\in N^L(k,l)} w(k,l,i,j,t) s_t^{Lc}[i,j]}{\sum_{t\in[1,\ldots,T]} \sum_{(i,j)\in N^L(k,l)} w(k,l,i,j,t)} \tag{5}$$

where the weights are

$$w(k,l,i,j) = e^{-\frac{||\hat{R}_{k,l} LL_t^{Lc}[k,l] - \hat{R}_{i,j} LL_t^{Lc}[i,j]||_2^2}{2\sigma_r^2}} \cdot f(\sqrt{(k-i)^2+(l-j)^2}) \tag{6}$$

where $\hat{R}_{k,l}$ is an operator that extracts the patch from image at location (k,l). We introduce the luminance values in search because they help finding adequate closest points because they contain patterns that facilitate search for the nearest neighbourhoods.

As in the case of reconstruction of low-frequency depth coefficients reconstruction, preliminary reconstruction of high-frequency coefficients can be carried out using coefficients from neighborhood. For the HL-subbands, a vertical linear interpolation is a good initial estimate of missing coefficient, since there exist mainly vertical correlation between coefficients due to horizontal high pass filtering. Similarly for the LH-subbands, horizontal linear interpolation creates good initial estimate. In the case of HH-band coefficients are replaced by zero, since most of these coefficients are zero. This replacements did not caused disturbing artefacts.

Initial interpolation of high frequency content is again refined through search for best matching blocks in the previous frames. This time for each neigborhood we form integral edge of both luminance and depth by taking an absolute value of element-wise product. This is done to identify the edges of object and at the same time to diminish the influence of textured parts. We extract 5x5 patches from all frames of depth and luminance images and form kd-tree structure from lexograficaly ordered pixels contained in patches. Then for each pixel of the high frequency coefficients, the closest patches in each frame are found. The patches of the integral images are formed to avoid the higher dimensionality of the kd-tree, which would reqire much more memory and calculations for its creation. For high frequency content restoration, we perform test to detect whether the large coefficient in the current patch originates from the edge or from texture in luminance frame. If the correlation between luminance and depth is higher than some empiricaly determined threshold, we use the edge pixels of luminance to interpolate the depth high frequency content, otherwise we use samples of low frequency depth. We use empirical threshold since the sequences that are used for the test are of high quality, and do not contain noise. We choose them in order to be able to quantitatively measure the performance of the proposed algorithm.

The value of high frequency wavelet coefficients of depth is obtained using Eq. 7

$$\hat{d}_d^h = \frac{\sum_{t\in[1,\ldots,T]} \sum_{(i,j)\in N^L(k,l)} w(k,l,i,j,t) d_t^{Lic}[i,j]}{\sum_{t\in[1,\ldots,T]} \sum_{(i,j)\in N^L(k,l)} w(k,l,i,j,t)} \tag{7}$$

where $\hat{d}_d^h$ stands for the estimated value of the wavelet coefficient of depth, $d_t^{Lic}[i,j]$ is the preliminary value of the interpolated pixel, and $w(k,l,i,j,t)$ is the normalized value of the influence of the pixel with coordinates (i,j) from the frame t on an interpolated pixel at location (k,l). We define interpolation weights as

$$w(k,l,i,j) = e^{-\frac{||\hat{R}_{k,l} HF_t^{Lc}[k,l] - \hat{R}_{i,j} HF_t^{Lc}[i,j]||_2^2}{2\sigma_r^2}} \cdot f(\sqrt{(k-i)^2+(l-j)^2}) \tag{8}$$

where $HF(i,j)$ is the value of the high frequency component at location (i,j).

5. EXPERIMENTAL RESULTS

We test the performance of the proposed algorithm on four sequences: "Interview", "Orbit", "Cg" and rendered sequence of a street with global motion and artificial textures. First two sequences represent real scenes. Resolution of all sequences is 720x576 and correspond to PAL standard resolution. To experimentaly validate our algorithm, we decimate depth sequences 8 times using db4 wavelets as anti-aliasing filters. This situation correspond to the depth field estimated using block disparity estimation algorithm e.g.[4] with the block size 8. Performance of the proposed method is compared to a depth map obtained using

nearest neighbor interpolation (equivalent to low resolution depth image), Lanczos inerpolation method[21] and superresolution method of Farsiu from[7] applied to the sequence of depth images and show significant improvements in subjective and PSNR term. Quantitative results are given in in Table 1. We have also made a comparison of artificial views generated using depth maps obtained using different methods. Virtual view generated by depth map obtained using the proposed method contains the least number of occlusions as shown in Figure 5. Example of superresolved depth image is shown in Figure. 4. From Figure. 4 we can see that the depth map was successfuly upsampled withou losing sharpness of the edges.

Table 1. PSNR values of superresolved depth images

Method	PSNR value
Low resolution depth image	23.53
Lanczos interpolation	24.6
Farsiu[7]	27.25
proposed	28.18

6. CONCLUSION

In this paper we present method for superresolution of depth images, based on using luminance images and non-local processing. Proposed method preserves depth image details, because it takes into account the most similar patches to the interpolated one. Method avoid implicit motion estimation since those methods can cause disturbing artefacts in the case of occluded sequence parts. In our future work we plan to use more sophisticated methods for searching similar contexts in n-dimensional space, to achieve more precise context modelling.

REFERENCES

1. J. Diebel and S. Thrun, "An application of markov random fields to range sensing," *Advances in Neural Information Processing Systems* (18), pp. 291–298, 2006.
2. A. Riemens, O. Gangwal, B. Barenbrug, and R.-P. Berretty, "Multi-step joint bilateral depth upsampling," *SPIE Visual Communications and Image Processing* **7257**, 2009.
3. S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "High-quality scanning using time-of-flight depth superresolution," *CVPR Workshop on Time-of-Flight Computer Vision 2008* , 2008.
4. G. de Haan, P. Biezen, H. Huijgen, and O. Ojo, "True motion estimation with 3-d recursive search block-matching," *IEEE Trans. on Circuits and Systems for Video Technology* **3**(5), pp. 368–388, 1993.
5. Q. Yang, R. Yang, J. Davis, and D. Nister, "Spatial-depth super resolution for range images," *IEEE Computer Vision and Pattern Recognition*, 2007.
6. O. P. Gangwal and R. P. Robert-Paul Berretty, "Depth map post-processing for 3d-tv," *IEEE International Conference*, pp. 231–242, 1998.
7. S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multi-frame super-resolution," *IEEE Transactions on Image Processing* **13**(10), pp. 1327–1344, 2004.
8. J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele, "An application of markov random fields to range sensing," *ACM Trans. on Graphics* **26**(3), 2007.

9. M. Lindner, M. Lambers, and A. Kolb, "Edge-enhanced distance refinement for 2d rgb and 3d range images," *International Journal of Intelligent Systems Technologies and Applications* (Issue on Dynamic 3D Imaging), 2008.
10. M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the non-local-means to super-resolution reconstruction," *IEEE Transactions on Image Processing* **18**(1), pp. 36–51, 2009.
11. C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *ICCV*, p. 839846, 1998.
12. E. Eisemann and F. Durand, "Flash photography enhancement via intrinsic relighting," *ACM Trans. on Graphics* **23**(3), pp. 670–675, 2004.
13. G. Petschnigg, M. Agrawala, H. Hoppe, R. Szeliski, M. Cohen, and K. Toyama, "Digital, photography with flash and no-flash image pairs," *ACM Trans. on Graphics* **23**(3), 2004.
14. A. Pižurica and W. Philips, "Estimating probability of presence of a signal of interest in multiresolution single- and multiband image denoising," *IEEE Trans. on Image Processing* **15**(3), pp. 654–665, 2006.
15. S. Chang, Z. Cvetkovic, and M. Vetterli, "Fast motion vector estimation using multiresolution-spatio-temporal correlations," *IEEE Trans. on Image Processing* **15**(6), pp. 1471–1485, 2006.
16. M. Zhang and B. Gunturk, "Multiresolution bilateral filtering for image denoising," *IEEE Trans. on Image Processing* **17**(12), pp. 2324 – 2333, 2008.
17. J. Canny, "A computational approach to edge detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **8**(6), pp. 679–698, 1986.
18. V. Velisavljevic, "Edge-preservation resolution enhancement with oriented wavelets," in *IEEE Int. Conf. on Image Proc. (ICIP)*, (San Diego, USA), 2008.
19. X. Li, "Image resolution enhancement via data-driven parametric models in the wavelet space," *EURASIP Journal on Image and Video Processing* , 2007.
20. J. Rombaut, A. Piurica, and W. Philips, "Passive error concealment for wavelet-coded i-frames with an inhomogeneous gaussmarkov random field model," *IEEE Transactions on Image Processing* **15**, pp. 783–796, 2007.
21. K. Turkowski and S. Gabriel, *Filters for Common Resampling Tasks. Graphics Gems I*, Academic Press, 1990.
22. D. T. Lee and C. K. Wong, "Worst-case analysis for region and partial region searches in multidimensional binary search trees and balanced quad trees," *Acta Informatica* **9**, p. 2329, 1977.

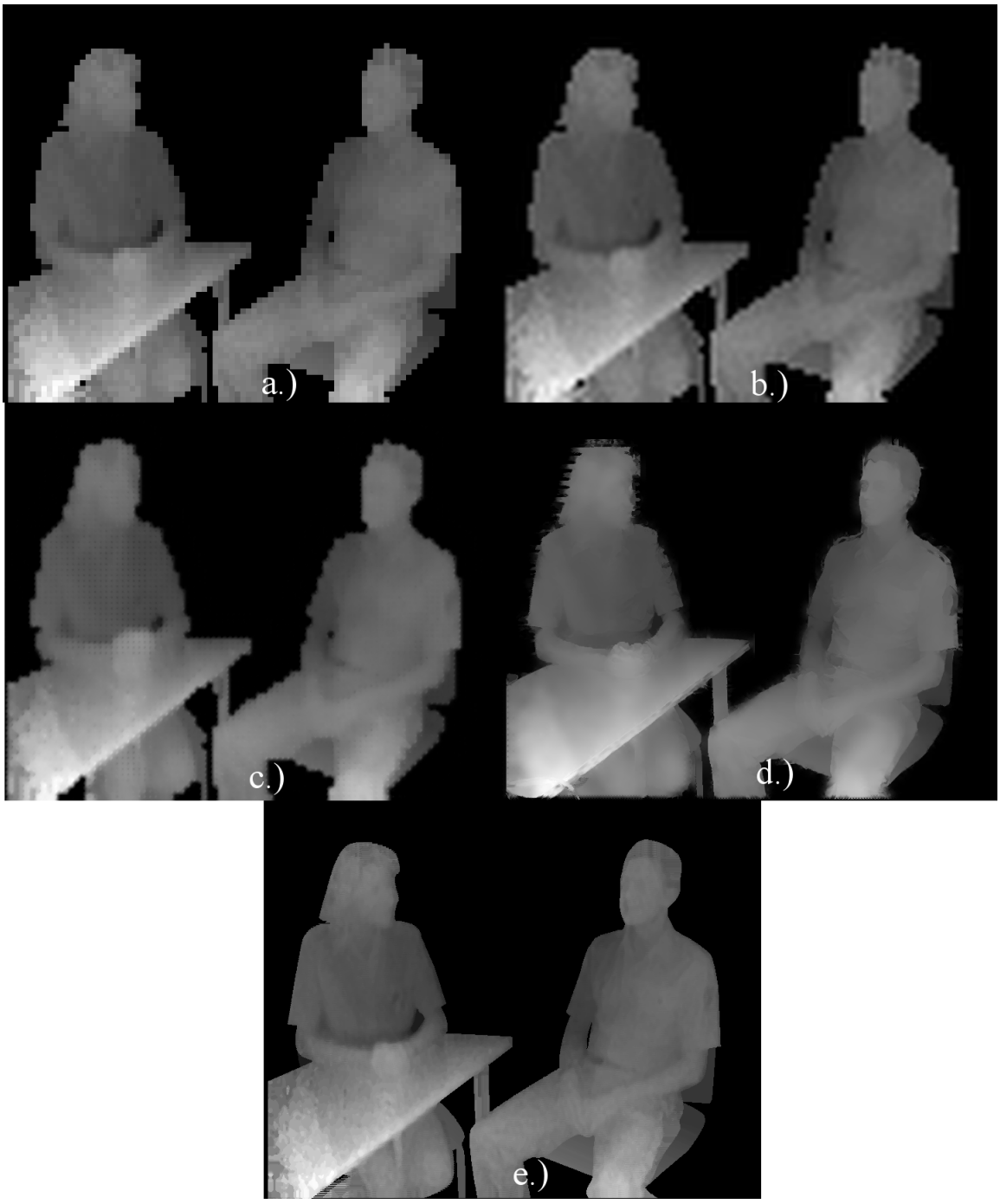

Figure 4. a.) Low resolution depth map b.) Depth map obtained by using Lanczos interpolation kernel c.) Depth map obtained using the superresolution method from Farsiu[7] d.) Depth map obtained by superresolution from proposed method e.) Original high resolution depth map

Figure 5. Artificial views generated using a.) Low resolution depth map b.) Depth map obtained by using Lanczos interpolation kernel c.) Depth map obtained using the superresolution method from Farsiu[7] d.) Depth map obtained by superresolution from proposed method e.) Original high resolution depth map