SUPER RESOLUTION OF TIME-OF-FLIGHT DEPTH IMAGES UNDER CONSIDERATION OF SPATIALLY VARYING NOISE VARIANCE

T. Edeler[1], K. Ohliger[1], S. Hussmann[1], and A. Mertins[2]

[1] Westcoast University of Applied Sciences (FHW), Heide, 25746 Germany
[2] Institute for Signal Processing, University of Lübeck, 23538 Germany

ABSTRACT

In this paper we propose a novel way of using time-of-flight camera depth and intensity images to produce a higher resolution depth image with prior knowledge of spatial noise distribution, which is correlated with the incident light falling on each pixel. The proposed method is compared to well-established methods, and results with real image data are presented.

Index Terms— Super Resolution, Time-Of-Flight, Denoising, Bilateral Filter, Regularization

1. INTRODUCTION

In recent years a new type of range sensors has conquered the market. Time-of-flight cameras provide, beside an ordinary 2D image, a depth map containing gray levels proportional to the distance of objects. Modern time-of-flight cameras have a spatial resolution of less than QVGA which is by far less than modern 2D CMOS/CCD image sensors. In addition to the poor resolution, the depth map provided by the camera is superimposed by a considerable amount of noise. In practice most applications average the depth maps of several subsequent images to increase the SNR. Fortunately the noise variance per depth pixel is correlated with the amount of light collected by that pixel.

In the literature two decades ago an interesting field in image processing emerged called super resolution, which has been extensively studied since then (see [1] and references therein). In this paper we show the combination of a denoising approach with the technique of super resolution for depth images. For this we exploit the fact that the gray-value image can be used to obtain an uncertainty map for the depth image and therefore can be used to improve denoising.

In the next section we introduce the image model, used for the theory of denoising and super resolution. Section 3 describes the important part of fusing the measured images to a single high-resolution denoised result and the task of regularization, since both denoising and super resolution are ill-posed problems. Section 4 gives an introduction to time-of-flight measurements and the underlying noise model. Finally the results are presented in Section 5.

2. GENERATIVE IMAGE MODEL

Let us assume we have several images taken from a real scene (continuous in space and intensity) by an image sensor through an optical system and underlying some spatially and time dependent geometric transformations (e. g. through camera movement) as well as noise. Further we assume that this real scene is limited in its spatial frequency and that there exists an exact representation in the discrete spatial domain [2]. This representation is the vector $\mathbf{x}$ of dimension $[LN_1 \cdot LN_2 \times 1]$. All images in this work are represented by column wise lexicographic ordered vectors. Each image taken by the camera system with a resolution of $N_1 \times N_2$ can be modeled as

$$\mathbf{y}_k = \mathbf{DHF}_k\mathbf{x} + \mathbf{v}_k, \tag{1}$$

where $\mathbf{D}$, $\mathbf{H}$, and $\mathbf{F}_k$ are matrices representing the downsampling, blurring and geometric transformation, respectively. $\mathbf{v}_k$ represents zero mean Gaussian noise with a covariance matrix of $\Sigma_k = \mathrm{diag}[\sigma_1^2, \sigma_2^2, ..., \sigma_{\mathrm{N_1N_2}}^2]$. The k-th camera image is represented by $\mathbf{y}_k$.

3. DATA FUSION

3.1. Single-image restauration under assumption of spatially variant noise variance

Let us assume we have taken one image $\mathbf{y}$ according to the model in (1) with matrices $\mathbf{D}$ and $\mathbf{F}_k$ being identity matrices. The image can then be written as

$$\mathbf{y} = \mathbf{Hx} + \mathbf{v} = \mathbf{z} + \mathbf{v}. \tag{2}$$

Instead of formulating the problem of finding $\mathbf{x}$, we try to find $\mathbf{z}$ first and therefore separate the reconstruction process into a denoising and deconvolution part [3]. Looking at this problem as a statistic optimization problem, our objective is to maximize $p_{\hat{\mathbf{z}}|\mathbf{y}}(\hat{\mathbf{z}}|\mathbf{y})$. According to the Bayes theorem

$$p_{\hat{\mathbf{z}}|\mathbf{y}}(\hat{\mathbf{z}}|\mathbf{y}) = p_{\mathbf{y}|\hat{\mathbf{z}}}(\mathbf{y}|\hat{\mathbf{z}})p_{\hat{\mathbf{z}}}(\hat{\mathbf{z}})\frac{1}{p_{\mathbf{y}}(\mathbf{y})}, \tag{3}$$

the objective can be seen as maximizing $p_{\mathbf{y}|\hat{\mathbf{z}}}(\mathbf{y}|\hat{\mathbf{z}})p_{\hat{\mathbf{z}}}(\hat{\mathbf{z}})$. Under the assumption of equal probability for all $\mathbf{z}$ our problem can be stated as

$$\hat{\mathbf{z}} = \arg\max_{\mathbf{z}} p_{\mathbf{y}|\mathbf{z}}(\mathbf{y}|\mathbf{z}), \tag{4}$$

which is the formulation of the maximum likelihood estimation. Given the model in (2) with $\mathbf{v}$ being white Gaussian noise the likelihood can be written as

$$p_{\mathbf{y}|\hat{\mathbf{z}}}(\mathbf{y}|\hat{\mathbf{z}}) = \prod_{i=1}^{N_1 \cdot N_2} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - \hat{z}_i)^2}{2\sigma_i^2}\right). \tag{5}$$

Consequently the log likelihood yields

$$\mathcal{L}(\hat{\mathbf{z}}; \mathbf{y}) \propto -\sum_{i=1}^{N_1 \cdot N_2} \left(\frac{(y_i - \hat{z}_i)^2}{\sigma_i^2}\right) = -(\mathbf{y} - \hat{\mathbf{z}})^T\mathbf{W}(\mathbf{y} - \hat{\mathbf{z}}). \tag{6}$$

with $\mathbf{W} = \mathrm{diag}[\sigma_1^2, \sigma_2^2, ..., \sigma_{\mathrm{N_1N_2}}^2]^{-1}$ containing the weights. Equations (2), (4), and (6) lead to a solution for the problem of estimating $\mathbf{x}$:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \Big[(\mathbf{y} - \mathbf{H}\mathbf{x})^T \mathbf{W} (\mathbf{y} - \mathbf{H}\mathbf{x}) \Big] \tag{7}$$

3.2. Super resolution data fusion

Let us assume we have several downsampled, blured, and geometrically transformed images $\mathbf{y}_k$. The goal of superresolution is to reconstruct the high-resolution image $\mathbf{x}$ from the observations $\mathbf{y}_k$. The full model (1) applies and (7) becomes

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \sum_{k=1}^{p} \big[(\mathbf{y}_k - \mathbf{DHF}_k\mathbf{x})^T \mathbf{W} (\mathbf{y}_k - \mathbf{DHF}_k\mathbf{x}) \big]. \tag{8}$$

In [3] it is shown that for a given pixel in $\hat{\mathbf{x}}$, (8) leads to the mean value of all pixels in $\mathbf{y}_k$ contributing to that single high-resolution-pixel.

3.3. Regularization

Since image reconstruction in general and super resolution in particular is an ill-posed problem [1], the solution of (8) will give in most cases a highly unwanted and unstable result. There is simply not enough information available to complete the task. This is when regularization comes into play. Its task is to incorporate prior knowledge of the desired solution in the minimization problem. This prior knowledge is usually expressed by a penalty term in the objective function:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \sum_{k=1}^{p} \big[(\mathbf{y}_k - \mathbf{DHF}_k\mathbf{x})^T \mathbf{W} (\mathbf{y}_k - \mathbf{DHF}_k\mathbf{x}) \big] + \lambda \Phi(\mathbf{x}) \tag{9}$$

where $\Phi(\cdot)$ penalizes unexpected solutions and factor λ weights the trade off between the measured data and the expected solution.

One of the most-widely referenced regularization cost functions is the Tikhonov ([4], [5]) cost function

$$\Phi(\mathbf{x}) = ||\Gamma\mathbf{x}||_2^2, \tag{10}$$

where Γ is a matrix incorporating some kind of filter. The intention is to limit the total energy in the image (when Γ is the identity matrix), or the Energy of a particular frequency band. For instance a high pass filter would lead to a smooth solution, since energy in the high-frequency band would be penalized.

Another very successful method in image reconstruction is the total variation (TV) method [6]. It penalizes the total amount of gray-level changes in an image, which is measured by the L_1 norm of the gradient:

$$\Phi(\mathbf{x}) = ||\nabla\mathbf{x}||_1. \tag{11}$$

The TV method has the advantage over other cost functions that it tends to preserve edges in the image, since steep edges do not count as much as they would in a quadratic cost function.

Based on the bilateral filter [7] and its link to the weighted least-squares method [8], Farsiu et al. proposed a robust regularizer called bilateral total variation (BTV) [3]. It combines the robustness of the L_1 norm against noise with the basic idea of the bilateral filter to decay filter coefficients not only with the geometric distance to the center, but also with the distance of the gray-levels to the center. The regularization term they use is

$$\Phi(\mathbf{x}) = \underbrace{\sum_{l=-P}^{P} \sum_{m=0}^{P}}_{l+m\geq 0} \alpha^{|m|+|l|} ||\mathbf{x} - S_x^l S_y^m \mathbf{x}||_1, \tag{12}$$

where the matrices S_x^l and S_y^m are shift operators in x and y direction by an amount of m and l pixel respectively. Therefore the difference in (12) represents different scales of derivatives of $\mathbf{x}$, and $\alpha \in [0, 1]$ is used to weight them.

Fig. 1 shows results for the previously mentioned methods applied to a noisy image. Clearly the BTV criterion gives the best results in terms of edge preservation and noise reduction.

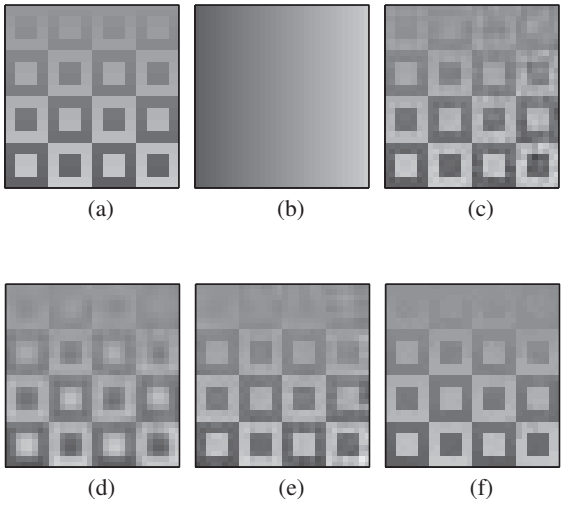

Fig. 1. Example of denoising with spatially variant variances. (a) Original Image. (b) Noisemap (black: $\sigma^2 = 0$ white: $\sigma^2 = 0.01$). (c) Noisy image. (d) Reconstruction with Tikhonov regularization. (e) Reconstruction using total variation. (f) Reconstruction using bilateral total variance.

4. TIME-OF-FLIGHT PRINCIPLE AND NOISE MODEL

Each time-of-flight camera is equipped with its own source of light. An object in a distance d from the camera (and its light source) reflects photons stemming from the modulated light source. They are collected by the time-of-flight pixel as

$$s(t) = a_0 \cos(\omega_0 t - \phi) + B, \tag{13}$$

where $s(t)$ is the average number of photons per unit time at given time t, ϕ is the phase shift resulting from photons traveling to the object and back to the camera $\left(\phi = 2\omega_0 \frac{d}{c}\right)$, with c being the speed of light. Thus the phase shift has a linear dependency from the distance of the reflecting object. The average incident light is taken into account by B.

Since the phase shift can not be measured directly, many time-of-flight systems use a pixel structure that performs some correlation

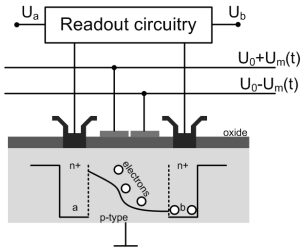

Fig. 2. Cross section of a single time-of-flight PMD pixel containing two wells.

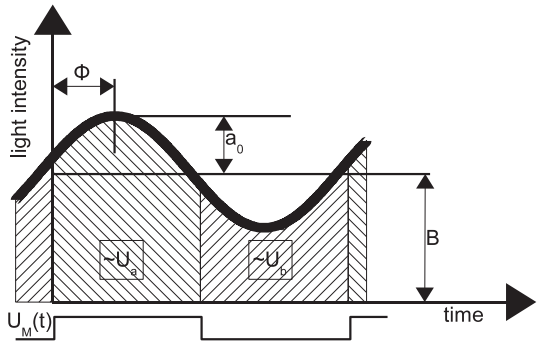

Fig. 3. Light intensity "seen" by a pixel over time. Correlation Voltage $U_m(t)$ is shifted by 0° to the modulation of the light source.

of the optical received signal with an electrical reference source. The pixel structure used for our experiments is shown in Fig. 2. The modulation signal $U_m(t)$ (see Fig. 3) directs electrons (caused by incoming photons) to either of two wells (a and b). To measure the phase shift of incoming light, four images (each taken with $U_m(t)$ shifted by 90° to its predecessor) have to be acquired. Fig. 3 shows the light intensity integrated by a pixel using $U_m(t)$ shifted by 0° to the emitted modulated light. Each pixel provides two voltages U_a and U_b. The differences ($\Delta U = U_a - U_b$), sampled at the four phase shifts, are used to calculate the modulation amplitude a_0 and phase shift ϕ of the optical echo [9]:

$$a_0 = \frac{1}{2}\sqrt{(\Delta U_{270} - \Delta U_{90})^2 + (\Delta U_0 - \Delta U_{180})^2} \qquad (14)$$

$$\phi = \arctan\left(\frac{\Delta U_{270} - \Delta U_{90}}{\Delta U_0 - \Delta U_{180}}\right). \qquad (15)$$

The number of photons collected during integration is underlying a Poisson distribution (even with perfectly constant intensity) [10]. In practice, when collecting several hundreds of photons, the distribution can be approximated by the normal distribution with the same value for mean and variance. This photon shot noise is responsible for the fact that a pixel collecting more light also outputs more noise (even though the SNR gets better). Since the phase shift of the optical echo does not depend on the total amount of light (but phase noise does), the phase SNR lowers when the nonmodulated light ($B - a_0$) gets brighter or the modulated light (a_0) gets darker. The dependency of phase noise is derived in [11]:

$$\sigma^2_{depth} \propto \frac{B}{a_0^2}, \qquad (16)$$

where σ^2_{depth} is the variance of the depth signal.

5. RESULTS

In this section the theory of the sections above is tested on real data. In our experiments we used the time-of-flight camera *PMD[vision] 19k*. All measurements were taken without any nonmodulated light. The test objects were a white colored styrofoam form and two forks mounted on an aluminum beam.

Taking equation (16) and setting $B = a_0$ (no nonmodulated light) the amplitude image should be proportional to the reciprocal of the depth-variance image, which can be seen in Fig. 4.

(a) (b) (c)

Fig. 4. Correlation between depth variance and amplitude image. (a) Single depth image. (b) Single amplitude image. (c) Reciprocal of variance image of 50 depth images.

With the theory of Sections 3 and 4 we optimize the super resolution and denoising for time-of-flight depth images by adding spatially dependent noise variance to the data-fidelity term in [3] (see (9)). In Fig. 5 our method of noise consideration is compared to the method of Fasiu et al. [3], which was also used in [12] to perform super resolution on depth images. In two experiments, four and 16 shifted low resolution images were taken and combined to an image with twice the resolution in both dimensions, see Fig. 5. Registration was performed using the algorithm in [13] on the amplitude images. Both experiments show that our method successfully combines the prior knowledge of variance distribution with the well studied SR method from [3]. The standard deviation is measured over three regions marked in Figs. 5(c) and (d). The results are presented in Table 1. The top region with low illumination shows a significant reduction of standard deviation for our method. Also in the middle and bottom region, where low noise is present in the measured depth images, the standard deviation of our method is still better.

Regions in which more noise is present due to poor illumination are reconstructed with stronger emphasis on the used regularization term, whereas in regions of good illumination the measured data are favored.

Table 1. Distance to object, measured with a tape measure, and standard deviation of regions marked in Figs. 5(c) and (d).

Region	distance [cm]	**method in [3]** stddev (c) [cm]	**proposed method** stddev (d) [cm]
top	122	1.463	0.343
middle	111	0.338	0.324
bottom	103	0.230	0.227

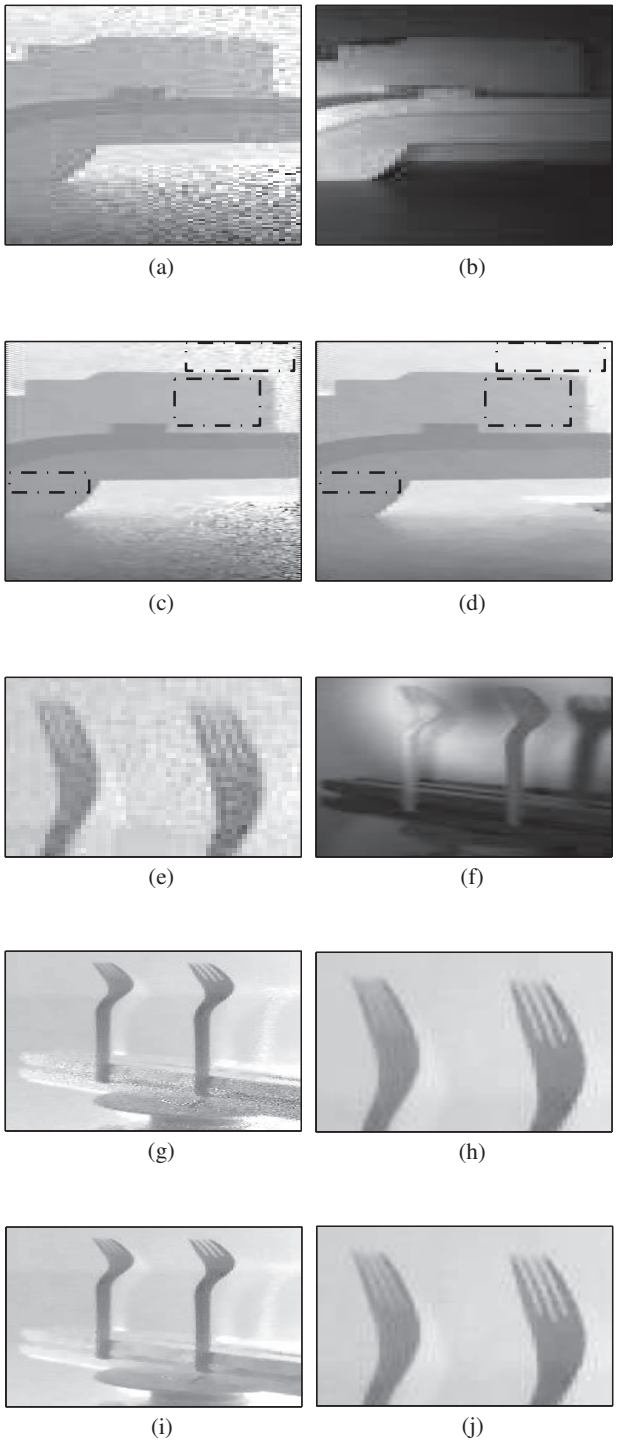

Fig. 5. Example of resolution enhancement and denoising of a depth image. Input depth images are used to gain a resolution enhancement by factor 2 in both directions. (a) One of four depth images. (b) Single amplitude image. (c) Reconstruction with the method in [3]. (d) Reconstruction of our method. – (e) One of 16 depth images (zoomed). (f) Single amplitude image. (g) Reconstruction with the method in [3]. (h) Zoom to fork spikes. (i) Reconstruction of our method. (j) Zoom to fork spikes for our method.

6. CONCLUSIONS

In this paper we introduced a new algorithm for enhancing resolution of time-of-flight depth images with respect to spatially variant noise variance. The noise intensity is estimated from the gray value image, which is always provided by the camera beside the depth image. We addressed different regularization terms for image denoising, where we chose to use the one with best results for typical depth images. For the purpose of comparison, evaluations on real-world time-of-flight images were made with a well established algorithm and our new one.

7. REFERENCES

[1] S. Borman and R. L. Stevenson, "Super-resolution from image sequences-a review," in *Proc. Midwest Symposium on Circuits and Systems.*, 1998, pp. 374–378.

[2] A. Papoulis, "Error analysis in sampling theory," in *Proc. IEEE*, 1966, vol. 54 of 7, pp. 947–955.

[3] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super resolution," *IEEE Trans. on Image Processing*, vol. 13, no. 10, pp. 1327–1344, 2004.

[4] M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Trans. on Image Processing*, vol. 6, no. 12, pp. 1646–1658, 1997.

[5] N. Nguyen, P. Milanfar, and G. Golub, "A computationally efficient superresolution image reconstruction algorithm," *IEEE Trans. on Image Processing*, vol. 10, no. 4, pp. 573–583, 2001.

[6] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, no. 1-4, pp. 259–268, 1992.

[7] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth International Conference on Computer Vision*, Stanford, USA, Jan 1998, pp. 839–846.

[8] M. Elad, "On the origin of the bilateral filter and ways to improve it," *IEEE Trans. on Image Processing*, vol. 11, no. 10, pp. 1141–1151, 2002.

[9] S. Hussmann and T. Liepert, "Three-dimensional tof robot vision system," *IEEE Trans. on Instrumentation and Measurement*, vol. 58, no. 1, pp. 141–146, Jan. 2009.

[10] J. Nakamura, *Image sensors and signal processing for digital still cameras*, CRC, 2006.

[11] R. Lange and P. Seitz, "Solid-state time-of-flight range camera," *IEEE Journal of Quantum Electronics*, vol. 37, no. 3, pp. 390–397, 2001.

[12] S. Schuon, C. Theobalt, J. Davis, and S. Thrun, "High-quality scanning using time-of-flight depth superresolution," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08.*, pp. 1–7, June 2008.

[13] S. Peleg, D. Keren, and L. Schweitzer, "Improving image resolution using subpixel motion," *Pattern Recognition Letters*, vol. 5, no. 3, pp. 223–226, 1987.