Joint de-noising and fusion of 2D video and depth map sequences sensed by low-powered tof range sensor

Citation

Year
2013

Version
Peer reviewed version (post-print)

Link to publication
TUTCRIS Portal (http://www.tut.fi/tutcris)

Published in
2013 IEEE International Conference on Multimedia and Expo Workshops, ICMEW 2013, San Jose, CA, USA, 15.-19.7.2013

DOI
10.1109/ICMEW.2013.6618264

Copyright
© 2013 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Take down policy
If you believe that this document breaches copyright, please contact tutcris@tut.fi, and we will remove access to the work immediately and investigate your claim.
JOINT DE-NOISING AND FUSION OF 2D VIDEO AND DEPTH MAP SEQUENCES SENSED BY LOW-POWERED TOF RANGE SENSOR

Mihail Georgiev¹, Atanas Gotchev¹, Miska Hannuksela²

Tampere University of Technology, Tampere, Finland¹; Nokia Research Center, Tampere, Finland²

ABSTRACT

We propose a joint de-noising and data fusion approach where the fused modalities come from conventional high-resolution photo or video camera and low-resolution range sensor of Time-of-flight (ToF) type, operating in restricted conditions of low-emitting power and low number of sensor elements. Our approach includes identifying the various noise sources and suggesting suitable remedies at particular stages of data sensing and fusion. More specifically, fixed pattern noise and system noise are treated at a preliminary denoising stage working on range data only. In contrast to other 2D video/depth fusion approaches, which suggest working in planar coordinates, our approach includes additional denoising refinement in the space of 3D world coordinate system (i.e. point cloud space). Furthermore, the high-resolution grid resampling is performed as an iterative non-uniform to uniform resampling based on the Richardson method. This improves the performance compared to approaches based on low-to-high grid upsampling and subsequent refinement. We report experimental results where the achieved quality of fused data is the same as if the ToF sensor was operating in normal (low-noise) sensing mode.

Index Terms— ToF, PMD, 2D/PMD, 2D/ToF, time-of-flight, denoising, artifacts, fixed pattern noise, point cloud

1. INTRODUCTION

Depth information in the form of depth map plays a substantial role in the emerging multimodal 3D video systems. It augments one or multiple views to form a multimodal representation of the 3D scene of interest instrumental for rendering desired virtual views. Capture and representation of depth maps has become an area of intensive research. One particular topic of interest is the way of direct capture of distances by active devices utilizing the so-called time-of-flight (ToF) principle [1]. In such devices, continuously-modulated harmonic signal is emitted and the distances are at camera sensor elements (e.g. CMOS or CCD) through computing the phase-delay between the emitted and reflected signals. ToF devices are examples of active depth sensing devices which return the result of depth (range) sensing in the form of matrix of pixels with intensities proportional to the corresponding distances. The spatial resolution of such sensors is rather low (e.g. 200x200 pixels) and usually they are used in hybrid system by combining them with one or multiple 2D cameras of usually higher resolution (e.g. 1080x1920 pixels)[2]. Such combination raises the problem of aligning the multimodal sensors and fusing the corresponding data preferably on the higher-resolution grid; a process, referred to as 2D/ToF data fusion. Computationally, this requires reprojection of the depth data from the low-resolution grid to world coordinates and then back on the higher-resolution grid followed by re-sampling to get the depth values at the grid points. Any depth measurement noise would dramatically deteriorate the quality of the fused data.

Calibration techniques for 2D/ToF setups [2] bear similarity to those in stereo-vision [3]. Range-specific calibration techniques have been proposed in [4].

The problem of denoising of ToF data has been addressed in a number of works [2, 5]. Modern denoising approaches, such as bilateral filtering [6] and non-local (patch based) filtering [7] have been modified to deal with ToF data [2, 8]. The problem of combining 2D camera with depth data has been considered as a pure depth upsampling problem [10] of pure depth denoising problem [2] or as a noise-aware upsampling [5].

One particularly under-studied problem is the problem of depth data denoising and upsampling when the range sensor is operating in poor imaging conditions [8]. In this work, we present a 2D/ToF data fusion approach combined with ToF data denoising applied on range data sensed in extremely poor conditions. More specifically, we suggest denoising techniques applied at particular stages of the data fusion chain and aimed at removing particular noise components. We also suggest an iterative upsampling procedure based on the Richardson method.

2. RGB/TOF FUSION MODEL

A general 2D/ToF fusion application includes a synchronizing module, which provides simultaneous capturing, storing and visualizing data of all camera sensors, and a software post-processing module which manipulates the captured data for mutual relation and representation.

The 2D/ToF calibration process is responsible for the estimation of camera capturing parameters that provides cross-modality camera relation for data projection [3].
parameters for ToF camera and 2D color camera include: corresponding focal lengths - $f_{ToF}$, $f_{2D}$, principal point coordinates - $(x, y)_{ToF}$, $(x', y')_{2D}$, pixel map coordinates - $(u, v)$, $(u', v')$, and those describing optical system distortions. The parameters defining relative camera shift and pose are given by translation vector - $B_{[3x1]}$ and rotation matrix $R_{[3x3]}$. When calibration parameters are estimated and optical distortions compensated, the process of data projection is straightforward. Each pixel of the ToF range map - $D(u, v)$ corresponds to world coordinates denoted as - $X$, $Y$, and $Z$ (depth): $[X, Y, Z]=D(u, v)/d(u, v)=[u-x_{ToF}, v-y_{ToF}, f_{ToF}]$, where $d(u, v)=\sqrt{(x_{ToF}-u)^2+(y_{ToF}-v)^2+f_{ToF}^2}$. We denote further this data representation as “point-cloud”. Then, a projection to pixel map coordinates of arbitrary 2D camera is given by the transform operation: $u_g, v_g, k = P[X, Y, Z, I]^T$, where $(u_g, v_g)$ are homogenous coordinates related to $(u', v')$ by $(u', v')=[u_g, v_g]/k$ and $P$ is projective matrix given by:

$$P = \begin{bmatrix}
 f_{2D} & 0 & p_2(x') \\
 0 & f_{2D} & p_2(y') \\
 0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
 R & B
\end{bmatrix}^T$$

(1)

The projected data has the nature of irregularly scattered samples and in case of low-resolution ToF device it is often very sparse density (c.f. Fig. 1). Relating and approximating the irregularly sampled data to denser regular coordinate grid positions of 2D camera is performed by a re-sampling and approximation process.

The visual result of 2D/ToF fusion output as depicted in Fig. 1 represents an ideal case of noise-free depth map. A high range error is expected when the ToF operating mode is forced to work in technologically limited hardware or when the device senses signals of very low amplitudes. A possible integration of ToF device in a portable device requires a re-engineered hardware solution. This includes lowering power of beaming unit, smaller number of LEDs, shorter integration intervals of sensor chips, cheaper hardware, etc. Such restricted conditions can be summarized as low-powered mode to differentiate from the normal operating mode. The former is more susceptible to measurement errors (range noise) manifested by moving fog of grainy particles on the sensed depth map. In our work, we are particularly interested in the ToF/ToF fusion process is affected by the influence of (extreme) range noise. The case is illustrated in Fig. 2. One can observe from there that while the original noisy range data has yet some distinguishable scene structure, the fused output is completely degraded and useless in practice. This illustrates the importance of proper care of the range noise across the 2D/ToF fusion steps so to avoid the amplification effect.

3. 2D/TOF FUSION MODEL IN THE PRESENCE OF STRONG NOISE

The first step in dealing with extreme noise in ToF measurement data is to remove the fixed pattern noise (FPN). It represents a spatially-related offset mask which appears as stripes of brighter or darker intensity across the image pixels. FPN removal is a rather trivial step not addressed in this paper. Then it is followed by ToF system noise removal, also applied on the range data only. The technique for tackling this type of noise has been thoroughly explained in our other work and attached as supplemental material [8]. The technique utilizes a non-local means denoising approach modified to work in a complex-variable domain [8]. In addition to removing system noise caused by the device working in low-powered mode, the proposed method was proved efficient for suppressing artifacts presented also in normal operating mode [8].

3.1. Surface mesh denoising of ToF data and 2D/ToF camera calibration

The method described in [8] allows for tunable edge preservation of the resulting denoised depth map for the price of some residual noise. Such noise can be further suppressed utilizing some additional structural information available in the sensed data. While the method in [8] makes use of the amplitude of the sensed signal as a confidence measure of the noise presented, our proposal is to make use of surface information presented in the point cloud of range measurements. Having these two denoising stages separated makes the whole process more controllable through involving simpler filters.

The point cloud of ToF data defines in fact a surface mesh (or waterbed mesh), which means that only a single point could exist on the optical ray formed by corresponding range pixel and camera center - Co. This property can be used for a surface mesh denoising in the flavor of the
techniques given in [9]. The main idea is explained as follows: for each vertex \( p = (X,Y,Z) \) of the surface mesh, define a noise-free surface \( S_{ll} \) by normal-unit vector - \( n_p \). The vertex \( p \) is updated along \( n_p \) as follows: \( p' = p + d(p)n_p \), where the amount \( d(p) \) is calculated by a bilateral filter [9]:

\[
d(p) = \frac{1}{C(p)} \sum_{n_p} \exp \left( - \frac{\|p - p\|^2}{2h_r^2} \right) \exp \left( - \frac{\|n_p - (p - p)\|^2}{2h_d^2} \right)
\]

where the inner product and vector norm give respectively the distance to noise-free surface and of surrounding vertices \( r \), \( C(p) \) is a filter normalizer, \( h_r, h_d \) are filter parameters, and \( R_V \) is a vertex ring of neighboring vertices \( r \). The vertices in \( R_V \) are selected as the closest vertices to the \( p \) denoted by \( R_{V1} \), or extended to closest vertices of those in \( R_{V2} \) denoted by \( R_{V2} \) or extended further in the same way for \( R_{Vn} \). Usually, \( S_{ll} \) is selected as normalized sum of normal-unit vectors of mesh polygons formed by the vertices of \( R_{Vn} \). However, we suggest that the noise-free surface \( S_{ll} \) for theToF surface mesh can be selected to be the one with normal-unit vector collinear to camera optical ray defined by corresponding range pixel to \( p \) and \( C_O \) (Figs. 3, 4), which follows from the given above constraint for point clouds of ToF data.

In contrast with the general data fusion chain where calibration takes place in the beginning, we suggest that the 2D/ToF calibration is to be performed only after the ToF system noise is filtered. This avoids spreading the measurement errors over the higher-resolution grid targeted. The correction of lens distortions should be combined with data projection to avoid additional re-sampling step. Resampling algorithms should be applied with care so not to distort data for occluded areas where no samples of projected data are available. Those areas should be excluded for resampling and filled in by some other algorithms [12]. Projected samples that are not visible in the rendered content are considered as “hidden points” and should be omitted as well. For occlusion detection we have applied the technique of shadow rendering as described in [13].

3.2. Approximation of projected irregular data

After surface refinement in point-cloud and occlusion filling, the depth data is reprojected in the coordinate system of the 2D color image. This reprojection results in samples irregularly positioned with respect to the given (high-resolution) grid which need to be upsampled at this grid. Our upsampling method makes use of the color information presented in the given 2D image, similarly to the denoising and upsampling approaches in [5, 10]. However, in contrast with these approaches, our method relies on an initial interpolation accomplished through careful fitting of Voronoi cells and subsequent iterations in the flavor of the Richardson method.

Consider a \( NxM \) discrete 2D color image in YUV color space denoted by \( y(n,m) = [y^r(n,m), y^g(n,m), y^b(n,m)] \), where \( n = 0,1,..,N-1; m = 0,1,..,M-1 \). The reprojected depth map is given at a set of irregular sampling points \( u'_i, v'_i, i=1,..,P \). We define a Voronoi interpolation operator \( V \) which brings the irregular samples to the regular grid of integers by fitting Voronoi cells around the given irregular samples [11]. The so-interpolated depth map \( z_{int}(n,m) \) undergoes a cross-bilateral filtering \( B \) making use of the available color information at the same nodes \( (n,m) \), thus obtaining the depth map at the \( k \)-th iteration:

\[
z_k(n,m) = r(n,m) \sum \frac{w_i}{\sum w_i} \left( \sum \frac{w_i}{\sum w_i} \left( (n-i)^2 + (m-j)^2 \right)^{1/2} \right) \left( y(n,m) - y(i,j) \right)
\]

where \( w_i = \exp(-x/y_x); a = s,c; \) and \( r(n,m) = \left( \sum \frac{w_i}{\sum w_i} \left( (n-i)^2 + (m-j)^2 \right)^{1/2} \right) \left( y(n,m) - y(i,j) \right) \)

The role of the cross-bilateral filter is to smooth any blocking artifacts arising from the nearest neighborhood Voronoi interpolation while preserving the object edges. The refined depth at the \( k \)-th iteration is then used to bilainearly interpolate the values at the starting irregular locations. Denote this bi-linear interpolator by \( L \). The procedure is then repeated on the error between the linearly interpolated depth values at the irregular positions and the initial depth values at the same irregular positions. Thus, the whole iterative procedure can be expressed as: \( z_{k+1} = B(z_k + \lambda V(z - Lz_k)) \), where \( \lambda \) is a tunable (relaxation) parameter. Fitting of Voronoi cells around given irregular samples is performed following the approach suggested in [11]. This improves the convergence. The use of bilinear interpolation for getting the depth values at the irregular locations is motivated by the observation that usually depth maps can be modeled as piecewise-linear functions at local neighborhood.

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

The experimental setup includes a 3D scene captured in varying sensing conditions. A precisely calibrated 2D/ToF
setup of ToF “PMDTech CamCube 2.0” and 2D “Prosilica GE 1900C” cameras has been used. Default ToF camera capture settings and sensor integration time of 2200 µs has been set to achieve normal sensing conditions. Such conditions usually ensure low noise level. Furthermore, 200 consecutively captured frames have been averaged given that the test scene is static. Ground true for the 2D/ToF fusion (GT2D/ToF) has been prepared by projecting the averaged ToF data on the high-definition grid and performing manual surface fit for each object, using markers. The varying test measurements are taken by four short ToF-sensor integration times, i.e. [500,200,100,50] µs.

### Table 1. Fusion results for different noisy conditions

<table>
<thead>
<tr>
<th>Method</th>
<th>Default</th>
<th>500 µs</th>
<th>200 µs</th>
<th>100 µs</th>
<th>50 µs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chan</td>
<td>44.12</td>
<td>39.64</td>
<td>32.54</td>
<td>30.43</td>
<td>23.41</td>
</tr>
<tr>
<td>Huhle</td>
<td>42.82</td>
<td>37.66</td>
<td>31.23</td>
<td>28.15</td>
<td>23.12</td>
</tr>
<tr>
<td>Proposed(NNN)</td>
<td>41.92</td>
<td>40.34</td>
<td>38.89</td>
<td>31.82</td>
<td>24.04</td>
</tr>
<tr>
<td>Proposed(1itr)</td>
<td>44.64</td>
<td>42.49</td>
<td>38.07</td>
<td>34.01</td>
<td>28.64</td>
</tr>
<tr>
<td>Proposed(2itr)</td>
<td>47.40</td>
<td>43.65</td>
<td>38.56</td>
<td>34.24</td>
<td>28.67</td>
</tr>
</tbody>
</table>

We include evaluation tests for our method and two state-of-the-art methods. Our method combines ToF system denoising with fusion based on iterative resampling approximation. The range data denoising is implemented as both planar and in point-cloud and the denoised data are upsampled and further refined by the Richardson procedure. The results are obtained for different number of iterations, denoted as “Proposed(1itr)”, and “Proposed (2itr)” respectively. The compared methods include the approaches by Huhle [2], and Chan [5] which apply fusion first and then denoise the depth data on the high-resolution grid. The results are benchmarked against GT2D/ToF in terms of PSNR and given in Table 1. Occluded areas have been excluded from the comparisons. Visual illustrations are provided in Fig. 6. The results are instructive about the superiority of the proposed joint denoising fusion approach in the case of high noise. Other denoising and upsampling methods seem to be effective only when applied in default operational mode, when the noise is relatively small. Our denoising approach followed by the proposed upsampling technique shows not only better results in terms of PSNR, but shows also very good visual quality comparable to the one obtained by ToF camera in default capturing mode (Fig. 7).

### REFERENCES


**Fig. 5.** ToF system noise denoising (top-to-bottom) noisy data, denoised data (left-to-right): a) 50 µs, b) 200 µs, c) 500 µs.

**Fig. 6.** Approximation results for (top-to-bottom) Chan, Huhle, Proposed(2-iter) (left-to-right): a) 500 µs, b) 200 µs, c) 50 µs.

**Fig. 7.** Output of joint-denoising fusion process for (left-to-right): a) GT, b) noisy (50 µs), and c) denoised.