TIME OF FLIGHT MOTION COMPENSATION REVISITED

J.M. Gottfried [*], *R. Nair*[*], *S. Meister*[*], *C.S. Garbe, D. Kondermann*

University of Heidelberg, Germany
Heidelberg Collaboratory of Image Processing

ABSTRACT

In this paper, we study motion artifacts that arise in Time-of-Flight imaging of dynamic scenes caused by the sequential nature of the raw image acquisition process used to compute the final depth image. Many methods for compensation of such errors have been proposed to date, but still lack a proper comparison. We bridge this gap by not only evaluating those methods, but also by providing implementations for all of them as a base-line to the community. By exchanging the calibration model necessary for these methods with a model closer to reality we were able to improve the results on all related methods without any loss of performance.

Index Terms— ToF, CWIM, motion-artifacts, calibration

1. INTRODUCTION

Today, Time-of-Flight (ToF) imaging is a mature technology, used in industrial applications like optical inspection, robot control and surveillance. Soon, with the new generation of low-cost sensors such as Microsoft's *Kinect 2* it will also hit the mass consumer market and therefore also impact human computer interaction, artistic expression and citizen sciences. Yet, like any other depth imaging modality, ToF data does have its own set of issues such as flying pixels, depth wiggling, multi-path and motion artifacts. In this paper we focus on motion artifacts and existing methods to deal with them. These artifacts occur in dynamic scenes due to the sequential nature of the measurement process and are a problem inherent to all current ToF cameras.

We consider *Continuous Wave Intensity Modulation* (CWIM) sensors as virtually all current ToF devices use this technique. They measure the phase shift between a reference sinusoid signal and the incident reflection of a light source modulated with the reference. This is done by sampling the correlation function between incident and reference signal for at least four phase shifts by recording four *raw frames*. Modern two-tap sensors[1] measure two of these raw frames during one single exposure (*subframe*). Yet, as these taps usually have different response curves [1], usually two more (redundant) subframes are recorded. Motion artifacts occur when the static scene assumption between subframes does not hold. Artifacts examples are depicted in fig. 1 (left).

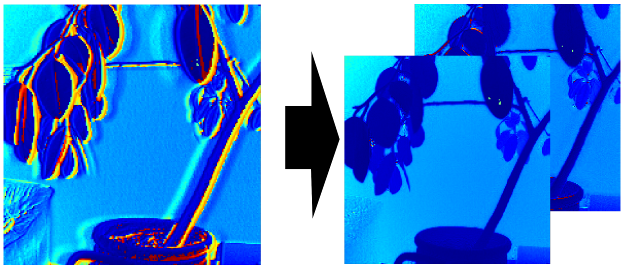

Fig. 1. Example depth map with motion artifacts (left) and results of presented methods (Schmidt's BID method (right, upper), Lindner&Kolb (right, lower)).

We offer three contributions: First, we provide the first comparison of this class of methods and offer all of them as open source implementations as a baseline for future research. Second, by identifying common building blocks used in many of the methods, we were able to implement them in a modular way. Using this framework we did not only benchmark the methods. We also evaluated the sensitivity of the processing pipeline towards the choice of different subcomponents (e.g the optical flow algorithm some systems use). Finally, during our investigations, we developed an improved cross-tap-calibration model. The usage of this new model leads to significantly improved motion compensation results in *all* proposed methods.

2. RELATED WORK

Existing methods for motion compensation can be classified into three categories. The first category contains methods that try to decrease the required number of sequential measurements. The next two categories both have in common that they incorporate a motion model into the reconstruction: Detect and repair methods *detect* regions affected by motion and *correct* them using local information. Flow based methods *estimate* the motion between subframes and *warp* the images according to the flow field before reconstruction. We now describe these methods in detail.

[*] The authors acknowledge financial support from the Intel Visual Computing Institute (IVCI, Saarbrücken).

[1] sensors with two light collecting units – called *tap* – per pixel

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

Framerate Increase by Tap Calibration: As mentioned before the measurements from two taps cannot be used together due to the different photo response. The relationship between the different taps is given implicitly per pixel by

$$A_i = r_i(B_i) \qquad i \in \{0, 90, 180, 270\} \qquad . \tag{1}$$

Schmidt [2] models these r_i as a linear polynomial and proposes a dynamic calibration scheme to estimate them.

With means of this calibration one may use the raw frames from only two exposures thus reducing, but not eliminating the effect of motion artifacts. This way, the eight raw frames are divided into two *subsets* S1 and S2.

Detect and Repair Methods: Such approaches can be further categorized in methods that operate directly on the depth image using additional inputs such as a foreground-background segmentation [3] or additional high resolution cameras [4] and the methods that harness the relation between the raw frames [2, 5, 6]. In our analysis we restrict ourselves to methods that do not require additional inputs. During the *detection* phase, Schmidt [2] uses the temporal derivatives of the individual raw frames. Motion artifacts occur if the first raw frame derivative is near zero (no change) whereas one of the other raw frames has a large derivative. Hansard *et al.* [6] operate on a similar principle, but evaluate the sums of two sub-frames. For the *correction* used to repair motion artifacts, Schmidt [2] uses the (temporally) last pixel values with valid raw images whereas Hansard *et al.* [6] the spatially nearest pixel with valid data.

Flow based Motion Compensation: Flow based methods [7, 8] loosen the requirement that the four measurements for reconstruction need to originate from the same pixel. Instead, the optical flow between subframes is used to find corresponding image locations. The application of optical flow to the raw data and the subsequent demodulation at different pixel positions require the following two points to be considered. *Brightness constancy*: optical flow methods often require corresponding surface points to have the same brightness. This is not the case for the raw frames. Fortunately, in two-tap sensors, photons are not "lost". The total amount of light in a pixel can be obtained by adding up the measurements of each subframe ($I_i = A_i + B_{i+180}$). Note that intensity subframes still do not satisfy brightness constancy as the primary light source still moves with the camera. Yet, in practice flow algorithms produce reasonable results on these images. The second point is *pixel homogeneity*: fitting the correlation function at different pixel locations requires a homogeneous sensor behavior over all locations. Again, this is not the case due to pixel gain differences and a radial light attenuation toward the image border. Lindner and Kolb [7] propose a raw value calibration based on work in [9]. The strength and weakness of this class of methods is strongly coupled with the flow method used. It is especially important to obtain the correct flow at occlusion boundaries, which at the same time pose a challenge to many flow algorithms.

3. PREPROCESSING AND CALIBRATION

We will first revisit the preprocessing steps needed by most of the algorithms, *i.e.* tap calibration and image homogenization. In our experiments, we worked with a CAMCUBE3 device by PMDTECHNOLOGIES.

For the subsequent methods, it is essential that the two record units (*taps*) of each pixel behave the same way. Especially the *framerate increase* would not work at all since it mixes values from both taps to estimate the phase shift of the returning light. Without any calibration one would take samples out of two different sine function (with different offset and amplitude) leading to arbitrary phase estimation results. So as a first step we have a look at the photo response of the taps of each pixel.

We recorded an *exposure ramp* of a flat white wall. The camera was mounted fixed to assure a static scene. Integration times were chosen in a range from the lowest possible value to saturation of the center pixels (0.1 ms up to 3 ms in steps of 0.1 ms). To avoid random noise, we took the average of 128 exposures with the same integration time.

Experiments comparing the tap A/B response for various pixels of the used camera show a non-linear behavior. For many pixels the graph has a strong curvature, at least for low intensities. Some pixels show similar behavior over the whole value range. This shows that the linear assumption of Schmidt [2] does not hold for the used device.

To overcome this issue, we propose a new reconstruction formula to cope with non-linear behavior at low intensities but linear extrapolation at high intensities. We used a polynomial fit of higher degree ($d = 5$ showed to be a good choice) and a linear fit for good extrapolation properties. The final reconstruction formula looks as follows:

$$r_c(b) = \Theta(b) \cdot r_1(b) + (1 - \Theta(b)) \cdot r_5(b) \tag{2}$$

with the polynomials r_d with different degrees d and an switching function Θ:

$$r_d(b) = \sum_{k=0}^{d} \alpha_k \cdot b^k \qquad \Theta(b) = \tfrac{1}{2}\left(\operatorname{erf}\left(\tfrac{b-\mu}{2\sigma}\right) + 1\right) \tag{3}$$

The switching function generates a smooth transition between the fitted linear and polynomial results and has the properties $\Theta(x \ll \mu) = 0$, $\Theta(x \gg \mu) = 1$. Since the central pixels get saturated at large integration times, we excluded high values from fitting. The parameters μ and σ are chosen to use r_5 for low and r_1 for high intensities as well as for extrapolation above the fit range.

Regarding the averaged phase using the raw pixel values of tap A and B together, application of the reconstruction r_c to the tap B values decreases the quality of the results as visible in fig. 2. Averaging the raw values of the two taps reduces the effects of artifacts of the individual taps whereas the reconstruction of tap B propagates the artifacts of tap A also to

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

the tap B values. So the reconstruction cancels some of the advantages of the averaging approach.

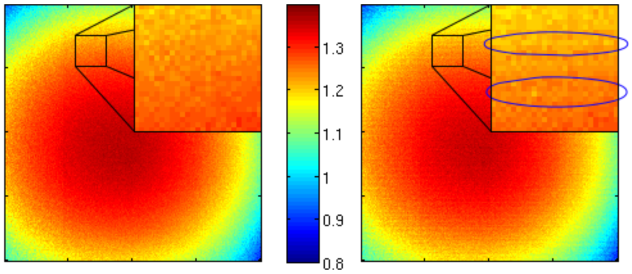

Fig. 2. Influence of the tap calibration on combined phase (averaging A and B). **left**: before, **right**: after application of r_c to tap B values. Zoomed region is magnified by four. Right picture shows line artifacts (marked by ellipses).

So tap calibration has advantages and disadvantages. Both taps of each pixel show different behavior which is helpful to average out pixel artifacts using all eight available uncalibrated raw data for phase estimation. So on static scenes, this averaging approach should be used without any reconstruction applied. At the other hand, decreasing the number of consecutive subframes to be captured to get all needed four phase shifts would decrease motion artifacts a lot. So assimilating the response of the taps is an important goal. Instead of picking one physical tap as a reference and assimilating the others one should use a ground truth photo response which would also tackle the problem of fixed-pattern-noise. But this ground truth is hard to estimate on the fly. Temperature and integration time dependency of the dark signal make this problem even harder and additionally impossible to calibrate this ground truth once and use it when record real sequences later. So the calibration approach presented here is the best one can do under the conditions of present ToF devices.

To be able to estimate the optical flow between the captured intensity images (sum of tap A and B of the individual sub-exposures), one has to use tap-calibrated data (after application of r_c to the tap B records) to avoid effects of different behavior of the taps. Second, there is a strong fixed pattern noise in the captured pictures as well as inhomogeneous lighting. Most optical flow methods do not work well under such conditions. Lindner and Kolb [10] propose a homogenization approach which is an *inter-pixel* calibration to circumvent these issues. Experiments showed that using the proposed form f

$$f(R_i) = a\sqrt{R_i + b} + cR_i + d \qquad g(R_i) = cR_i + d \qquad (4)$$

performs equally well as a simple linear formulation g (setting a, b to zero). R_i are the captured raw frames. The latter has the additional advantage that it does not change the value of the calculated phase since the parameters cancel out in the phase calculation formula and thus may be considered as being orthogonal to the *inter-tap* calibration presented before.

4. EVALUATION OF MOTION COMPENSATION

We evaluated all methods based on a real-world test sequence. We recorded a short *office sequence* of 256 frames. The camera was heavily rotated to obtain strong motion artifacts. The sequence contains diffuse as well as transparent and specular reflecting objects which are all within the non-ambiguity range. The camera position itself was almost kept constant so motion is only in the lateral and not in the depth direction.

Framerate Increase: Using the tap calibration presented before, it is possible to compute the phase of the reflected light with less than the four acquired exposures. The estimated phase results differ significantly depending on which data have been used to compute it. Using uncalibrated data, there are three options: All four exposures taken consecutively are needed. Only taking one tap for phase estimation shows strong artifacts, averaging them decreases the effects but fine structures have still up to four halo-like artifacts. As mentioned before, using corrected tap B data and averaging it with tap A decreases the quality benefit provided by the average. But using calibrated Tap B data, it is possible to compute the phase using the raw frames taken during the first two (S1) or the last two acquisitions (S2). Since there is only one delay between the expositions (instead of three delays using the standard approaches), this reduces the motion artifacts by a factor of three. Results clearly outperform the phase estimation methods using all four subframes (fig. 3 vs. 4, left), only a doubling effect at edges and fine structures remains.

Detect and Repair Methods by Schmidt: Fig. 3 shows the results of the standard motion compensation approach by Schmidt [2]. For simplicity, only the tap A phase images have been printed. Applied to the dynamic office scene, the results look really poor, the method fails completely on this sequence. The algorithm assumes that at most one discontinuity (called *event* in [2]) within two consecutive frames. This assumption is heavily violated in this real-world sequence. It may be fulfilled using a mounted camera in front of a almost

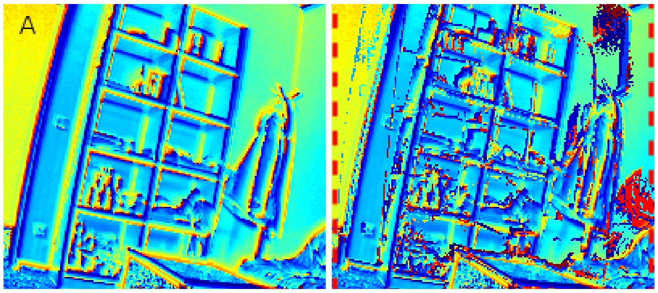

Fig. 3. Motion compensation method by Schmidt (Standard approach). Computed phases using all four tap A raw frames. **left:** before corretion. **right:** after correction. This method fails and causes strong artifacts at the considered sequence. It is not suitable for moving cameras.

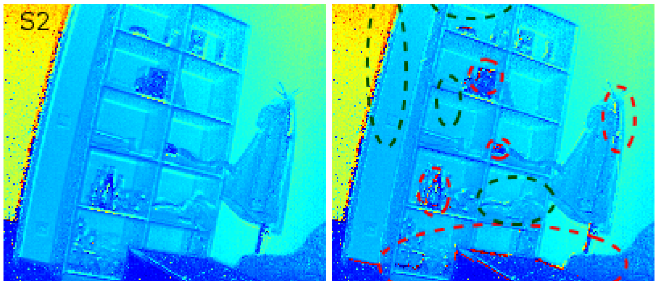

Fig. 4. Motion compensation method by Schmidt (BID approach). Phases computed using subset S2. **left:** before correction. **right:** with BID correction and marks of good/bad (green/red) results. The horizontal boards are sharper and with less halo-like artifacts. New artifacts appear, especially at the bottles but it works much better than standard approach.

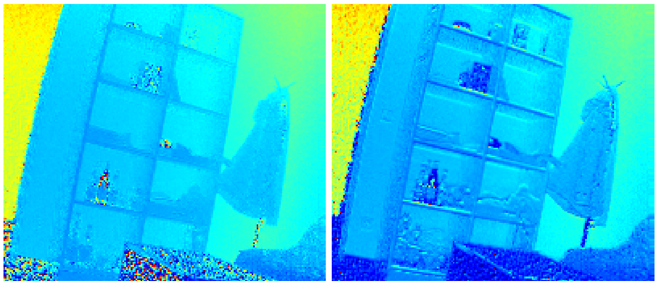

Fig. 5. Phase Visualizations after warping the raw frames (r_c corrected) with computed optical flow. Phase using all raw frames (averaging, **left**), and subset S2 (**right**), *i.e.*a combination of the Lindner and Kolb method with Schmidt's framerate increase. This gives best results but fails at low intensities.

static scene with only a few objects moving in lateral direction *e.g.*, object inspection at a conveyor belt.

The main problem using Schmidt's standard approach is that there are large displacements between two frames of the dynamic sequence violating the assumption of a most one discontinuity within two frames. The improved method called *burst internal detection* (**BID**) by Schmidt [2] does not use two frames but two subsets of one frame to detect and fix the discontinuities. This allows one event per frame without violating the assumptions. It also decreases the considered time window since the delay between all consecutive exposures to capture the subframes is much shorter than the delay between frames (in our case, all four subframes are captured in about $\frac{1}{4}$ of the time between two frames). Applied to our dynamic *office sequence*, the BID method gives the results shown in fig. 4. This is clearly an improvement compared to the standard method. There are still regions with artifacts, especially on translucent or reflecting objects but the doubling artifact occurring at depth edges in the uncorrected images is removed in most cases.

Optical Flow to Warp the Subframes: Since the presented system is intended to work in real-time, we focused on reference implementations of optical flow methods working on GPU devices (*i.e.* on graphic cards). Recently, many state-of-the-art methods [11, 12, 13, 14] have been implemented using Cuda and OpenCL in the OpenCV library. To simplify testing the different algorithms, all GPU methods provided by this library have been wrapped into Charon-Suite modules [15, 16]. This way, the common pre- and postprocessing parts could be performed using existing code. Experiments show, that optical flow results vary heavily depending on the used method. In our case, the TV-L^1 method [12] worked best. Fig. 5 shows the computed phase images using the raw frames warped with the TV-L^1 flow results. As interesting observations one should note that the quality strongly depends on the choice of phase computation method. Selecting all data from one single tap gives bad results. Especially the object edges show misalignment effects that look like aliasing. Combining images from both taps removes this kind of artifacts. Averaging all tap A and B exposures gives sharp edges and low noise in homogeneous regions. Only parts where the intensity was low (floor and box in the front) show high noise as well as the reflecting/translucent objects in the shelf. For some reason the noise in the low-intensity parts seems to be much higher in the first two exposures causing the depth maps computed using the first subset (bottom center) to be nearly useless there. Using the second subset only, results look drastically better. Edges are much sharper, even the bottles and dark objects in the shelf have been estimated correctly. Only at the horizontal wooden boards there are still some small halo-like artifacts that are not visible in the averaged image.

5. CONCLUSION

In this paper, we provide implementations and analysis of all purpose methods to cope with motion artifacts in time-of-flight images. A central part is the *tap calibration* proposed by Schmidt which has been extended to work with the non-linear behavior of the present capturing device. This calibration allows to use only two out of four exposures per frame reducing motion artifacts by a factor of three (*framerate increase*). This method leads to quite well results even on strong motion and should be sufficient for a large class of applications. If the remaining artifacts are still too strong, it should be combined with one of the subsequent methods. The BID method (detect and repair approach) by Schmidt uses the tap corrected data and replaces discontinuities in the second subset by values from the first one. This is a cheap and fast method but fails if the camera is moved heavily. More expensive but better results are given applying the flow method by Lindner and Kolb to the subsets. Using such framerate-enhanced data causes only small displacements which simplify optical flow computation and gives good results. Additionally only one flow field has to be computed instead of three (as in the traditional approach using all four exposures).

6. REFERENCES

[1] Michael Erz, *Charakterisierung von Laufzeit-Kamera-Systemen für Lumineszenz-Lebensdauer-Messungen*, Dissertation, IWR, Fakultät für Physik und Astronomie, Univ. Heidelberg, 2011.

[2] Mirko Schmidt, *Analysis, Modeling and Dynamic Optimization of 3D Time-of-Flight Imaging Systems*, Dissertation, IWR, Fakultät für Physik und Astronomie, Univ. Heidelberg, 2011.

[3] Salih Burak Göktürk, Hakan Yalcin, and Cyrus Bamji, "A time-of-flight depth sensor-system description, issues and solutions," in *Computer Vision and Pattern Recognition Workshop (CVPRW'04)*. IEEE, 2004, vol. 3, p. 35.

[4] Oliver Lottner, Arnd Sluiter, Klaus Hartmann, and Wolfgang Weihs, "Movement artefacts in range images of time-of-flight cameras," in *International Symposium on Signals, Circuits and Systems, 2007. ISSCS 2007*. IEEE, 2007, vol. 1, pp. 1–4.

[5] Stephan Hussmann, Alexander Hermanski, and Torsten Edeler, "Real-time motion artifact suppression in tof camera systems," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 5, pp. 1682–1690, 2011.

[6] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu P. Horaud, *Time of Flight Cameras: Principles, Methods, and Applications*, SpringerBriefs in Computer Science. Springer, November 2012.

[7] Marvin Lindner and Andreas Kolb, "Compensation of motion artifacts for time-of-flight cameras," in *Dynamic 3D Imaging*, Andreas Kolb and Reinhard Koch, Eds., vol. 5742 of *Lecture Notes in Computer Science*, pp. 16–27. Springer Berlin Heidelberg, 2009.

[8] D. Lefloch, T. Hoegg, and A. Kolb, "Real-time motion artifacts compensation of tof sensors data on gpu," in *Proc. SPIE, Three-Dimensional Imaging, Visualization, and Display*. 2013, vol. 8738, SPIE.

[9] Michael Stürmer, Jochen Penne, and Joachim Hornegger, "Standardization of intensity-values acquired by time-of-flight-cameras," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08*. IEEE, 2008, pp. 1–6.

[10] Marvin Lindner and Andreas Kolb, "Compensation of motion artifacts for time-of-flight cameras," in *Dyn3D*, Andreas Kolb and Reinhard Koch, Eds. 2009, vol. 5742 of *Lecture Notes in Computer Science*, pp. 16–27, Springer.

[11] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV (4)*, Tomás Pajdla and Jiri Matas, Eds. 2004, vol. 3024 of *Lecture Notes in Computer Science*, pp. 25–36, Springer.

[12] Christopher Zach, Thomas Pock, and Horst Bischof, "A duality based approach for realtime tv-l^1 optical flow," in *DAGM-Symposium*, Fred A. Hamprecht, Christoph Schnörr, and Bernd Jähne, Eds. 2007, vol. 4713 of *Lecture Notes in Computer Science*, pp. 214–223, Springer.

[13] Gunnar Farnebäck, "Fast and accurate motion estimation using orientation tensors and parametric motion models," in *ICPR*, 2000, pp. 1135–1139.

[14] Bruce D. Lucas and Takeo Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI*, Patrick J. Hayes, Ed. 1981, pp. 674–679, William Kaufmann.

[15] Jens-Malte Gottfried and Daniel Kondermann, "Charon suite software framework," *IPOL 2012 Meeting on Image Processing Libraries*, 2012.

[16] "Charon suite project page," available online at: `http://charon-suite.sourceforge.net`.

978-1-4799-5751-4/14/$31.00 ©2014 IEEE