PHYSCON 2017, Florence, Italy, 17–19 July, 2017

DEPTH RANGE ADAPTATION TO VARIABLE SCALE IN 3D-SCENARIOS FOR DENSE SLAM

Emanuel Trabes[1,2] Mario Alberto Jordan[1,2,3]

[1]Argentinean Institute of Oceanography (IADO-CCT-CONICET).
Florida km.8, 8000, Bahía Blanca, ARGENTINA.

[2]Dto. Ingeniería Eléctrica y de Computadoras- Univ. Nacional del Sur (DIEC-UNS)
Av. Alem 1253, 8000, Bahía Blanca, ARGENTINA.

***Abstract*—This work deals with the problem of inadequacy of a fixed depth range in cost-volume-based dense methods for visual SLAM to fit different scales of the scenery. Mainly, the rapid changes of scale in the focused scene during navigation makes it necessary the adaptation of the depth range in order to maximize the resolution. We present a novel approach based on the distribution of the mass probability of the inverse depth estimations. We track continuously the wrapping of this function, keyframe by keyframe, and create corrections on the basis of soft restrictions to be effective in the next step. We illustrate the efficacy of the real-time approach employing a dataset of a scenario indoors.**

Keywords: Monocular SLAM, Direct dense method, Cost volume, Inverse depth map, Depth range.

I. INTRODUCTION

SLAM techniques are increasingly being employed in many areas of the Robotics, in aerial as well as in ground, mobile robot navigation, space, swarm and underwater applications [1].

Recently, direct dense and semidense methods employing photometric knowledge have become popular in SLAM, see for instance DTAM [3] and LSD-SLAM [4].

In particular DTAM makes use of a cost volume to gather photometric information for checking brightness-consistency between every pixel in a keyframe with respect to pixels on the respective epipolar lines of ancillary frames. Thereby, a global time-demanding cost optimization of the photometric errors is performed along numerous epipolar directions to finally outcome a depth estimation [3]. However, new incoming technological breakthrough like GPGPU commodities allow to implement the complex procedure of optimization avoiding overhead for ground-truth applications [5].

A inherent feature of the cost volume is the discretization of the inverse depth range with more resolution for short- than long-sight distances between camera an object. Besides, the inverse-depth setup taps efficiently the great parallax for short- and middle distances of objects to the camera. Nevertheless, since the amount of depth units in the depth discretization influences exponentially the time taken to optimize, this is customarily selected as relatively small value, say 32 or 64. This in turn will produce a stair-casing effect in the depth map surface although it is negligible for short and tolerable for middle distances when the regularization strength is well set up.

In complex indoor and outdoor scenarios, equally natural or man-made constructions are geometrically characterized by a wide range of dimensions. If the depth range spans a very low minimal and a very high maximal depths during the navigation, this surely will carry a significant drop in the resolution when all objects in the footage are for instance at middle distances.

In this paper we will place the resolution problem at the core of the analysis in relation to a cost-volume-based depth estimation. We provide a powerful yet simple approach to overcome the problem by conferring the map the maximal resolution which is possible for the given discretization. After describing the heuristic and stages of the design, we will illustrate the approach performance by way of a real dataset.

[3] Corresponding Author: Mario Alberto JORDAN. E-mail: mjordan@criba.edu.ar. Address: CCT-CONICET-IADO, Florida km.7, B8000FWB, Bahía Blanca, ARGENTINA

II. Problem Statement

We begin to illustrate the problem to be dealt with by depicting a cost volume with a scenario like of the Fig. 1 (*cf.* [3]). Therein three inert objects are captured by the camera many times from different point of views.

In this configuration, object A and object C (at least partially) should not be included in the depth map even though they are perceived from all point of views during the camera displacement. Object B on the contrary, can be detected seamlessly with all details perceived from all viewing angles provided by the keyframe and its ancillary frames.

In such cases, objects will be out of scale and the depth estimations yield incorrect. A correction of the limiting depths $\xi_{\min}$ and $\xi_{\max}$ has to be attained and this should at best be done from time to time.

The core of the problem is that a continuous adaptation of the depth extremes will require of knowledge of the scene depth, for instance through the estimates, but on the other side these need just the information of the depth range to perform the estimation. It seems to be a chicken-and-egg like problem.

So the range adaptation seems at first glance to be an unsolvable problem from a theoretic point of view.

III. Preliminaries

By way of Fig. 1 we can describe the depth estimator appropriately and afterwards be able to develop a suitable heuristic for a solution to the problem stated previously.

Assuming constant brightness of the scene in all positions of the camera, an energy cost functional is defined on the cost volume for determining the depths of all pixels in the keyframe.

For inverse depth dense mapping estimation, DTAM employs a projective photometric cost volume C_r from any number m of ancillary frames with short-baseline one to each other nearby, which are overlapped with a reference image named keyframe r. The set of these frames is referred to as $\mathcal{I}(r)$. Moreover, the per-pixel inverse depth map is $D_r : \Omega \rightarrow \overline{d} \subset R^{+} \cup \{0\}$ where $\overline{d}$ is a inverse discrete interval $[d_{\min}, d_{\max}] = [1/\xi_{\max}, 1/\xi_{\min}]$ with inverse depth elements d defined by an uniform discretization of $\overline{d}$ in q units (*e.g.*, q=64), and $\Omega \in R^2$ is the set of normalized pixel coordinates that includes the intrinsic camera calibration. The average photometric error for every pixel u_r of the RGB reference image I_r is computed by

$$C_r(\mathbf{u}, d) = 1/\left|\mathcal{I}(r)\right| \sum_{m \in \mathcal{I}(r)} \left\|\rho_r(I_m, \mathbf{u}, d)\right\|_1, \quad (1)$$

where I_m is one image of $\mathcal{I}(r)$, $\mathbf{u}$ a row of entries of C_r (geometrically a frustum of C_r), ρ_r the single pixel photometric error function of the distance d for a given $\mathbf{u}$ of an I_m and $\|.\|_1$ is the $\mathcal{L}_1$ norm.

The photometric error function ρ_r (termed PEF from now on) is defined as the difference of a keyframe pixel and every pixel of the corresponding epipolar line pertaining to I_m. All PEFs for $m \in \mathcal{I}(r)$) corresponding to one keyframe pixel are stacked in $\mathbf{u}$ in the cost volume and then averaged in the direction of $\mathbf{u}$ to yield C_r in (1)**.** Both a PEF as well as the averaged PEFs are not convex in the direction of $\mathbf{u}$, *i.e.*, they may have many local extremes in $[1/\xi_{\max}, 1/\xi_{\min}]$.

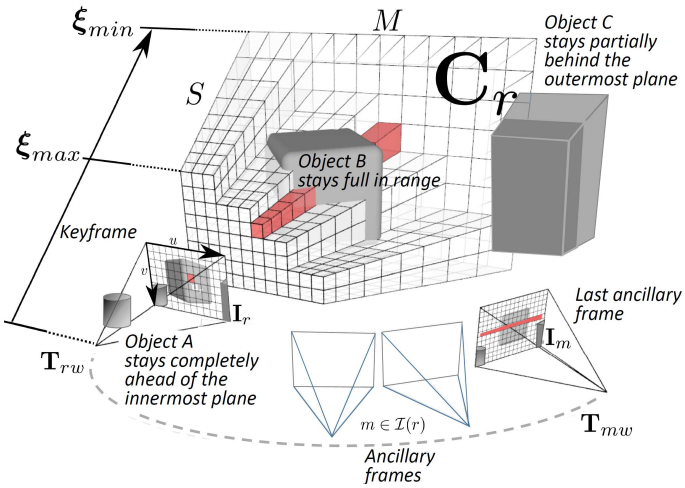

Figure 1 - Scene with 3 objects and cost volume (*cf.* [3]). Object A is outside depth range. Object B is in range. Object C is partially in range

Generally, it is expected that a physical point that is registered in the keyframe with coordinates (x, y) and also in many ancillary frames, satisfies

$$\min_{d \in [\xi_{\min}, \xi_{\max}]} C_r(\mathbf{u}, d) = \overline{C}_r(\mathbf{u}, \widehat{d}) \quad (2)$$

where $\overline{C}_r$ is usually a non-zero residual of the estimation and $\widehat{d}$ is unique and represents the depth of the physical point captured on the keyframe pixel (x, y). Owing to errors, mostly stochastic by nature, $\widehat{d}$ may be not unique or it might not exist or even in the absence of stochastic errors it might be for instance double $\widehat{d}$ because there are two physical points with the same brightness on the epipolar line.

Frequently, incoming blurred pictures of the scene give rise to many inconsistencies about the uniqueness of minima. In the worst case, this will ensue indetermination of C_r or generally untrue irregularity of the depth function $\widehat{d}(x, y)$ in a neighborhood of the keyframe pixel

(x, y). It is noticing that besides the noise interference over $\widehat{d}(x, y)$ there is a natural stair-casing effect produce by the discretization.

In order to overcome this problem and provide a more smooth but trustworthy estimations in the vicinity of every pixel and preserving natural edges, the cost function (1) is replaced by a non-convex energy functional [2]. This contains a non-convex photometric error data term and a convex spatial regularizer term integrating over all pixels that ensures a spatially smooth inverse depth map solution.

IV. Out-of-range depth estimation

It is clear that the approach for estimation described above can satisfactorily work with a setting equal to $[\xi_{\min}, \xi_{\max}] = [\varepsilon_0, \infty]$, where ε_0 is a positive real value close to zero and $\xi_{\max} = \infty$, *i.e.*, $d_{\min} = 0$ and $d_{\max} = \varepsilon_0^{-1}$.

Clearly this infinite span is appropriate to measure all the object depths, from a very close position to the camera, *i.e.* ε_0 in meters, up to objects in the far horizon. However the resolution achieved will be poor due to the stair-casing effect and it is more noticeable when the world objects are at equal distance from the camera.

Thereby, it is indispensable an adaptation of the depth range to the scene scale in order to capture fine details as a zoom function. For each physical point, which is common to many frames, there is an epipolar line by every one ancillary frame. This conforms a set of PEFs $C_r(\mathbf{u}_{x,y}, d)$ for every keyframe pixel (x, y),where $\mathbf{u}_{x,y}$ is the row in the cost volume.

The approach mechanism searches for photoconsistency, it is, equal pixel brightness on every epipolar line. As all objects are continuously focused by different camera poses, the mechanism will eventually assign a depth value within this interval independently if the object position is in or out of range.

Therefore, no matter how narrow is the span of $[d_{\min}, d_{\max}]$, the depth estimator mechanism will provide outcomes for all physical points captured in the sequence of frames, included for points outside the depth range. This fact is a key part of the heuristics developed to achieve our end.

V. Heuristic

Considering the depth value occurrence as a real-valued stochastic variable, this can be associated to a density function, which is defined over the discrete depth interval. This can be assessed through the histogram which is characterized by $f(\widehat{d})$. Certainly this is a stationary process when the scene topography is self-similar, but this is not the general general.

By the initialization of the depth map we will assume that all physical points of the scene are in range. It is clear that the related histogram on say $D_0 = [\underline{d}_0, \overline{d}_0] = [0, 1/\varepsilon_0]$ for ε_0 small satisfies this condition.

A property of the keyframe sequence in DTAM is that any two consecutive elements overlay a large portion of the scene. Hence, we can argue that their histograms will differ slightly. This feature points out that the probability mass function changes dynamically but to some extent slowly.

Thereupon we can change the next range interval D_1 by $[\underline{d}_0 + \underline{\delta}_1, \overline{d}_0 + \overline{\delta}_1]$ where $\underline{\delta}_1$ and $\overline{\delta}_1$ are real-valued corrections that enable the interval to stretch, compress or displace.

In so doing repeatedly with $D_2 ... D_t$ we are attempting to suitably put frame by frame the probability mass in the short-as-possible correct depth interval. As a result of this, the depth resolution is maximized for each keyframe and at the same time it is enabled that all scene physical points are included in the depth range correctly.

The heuristic suggests also that one must be aware that the physical points are not out of range, because in this case the depth estimator is not able to realize whether such points are outside the range. Insofar, the mass of probability must not cross the limits of the range interval but also it must not accumulate at its extremes.

VI. Adaptive resolution setting

As we are constrained to include permanently all point distances to the camera in a fluctuating depth interval and the span is also expected to be as short as possible, the rate of corrections has to be accomplished cautiously.

A. Distribution of probability mass

In Fig. 2. we illustrate the probability mass function at a given time point for three different scenes with one particular setting for each of them.

The function f_1 refers to a setting whereby some uncertain depth estimations in the keyframe have probably befallen. Therefore, the accuracy of some depth estimates in the keyframe is doubtful. Worse still, there is no information regarding which estimations are incorrect.

The function f_2 describes an optimal setting whereby the probability mass yields scattered on the whole interval with f_2 being null just at the interval extremes. Clearly, the resolution of the scene achieved by the estimation with this setting is in broad outlines high.

Finally, function f_3 exhibits a biased occurrence in the interval, which is a sign of an inadequate setting that

carries to loss of scene details, that it is to say, to get a low resolution of the focused scene.

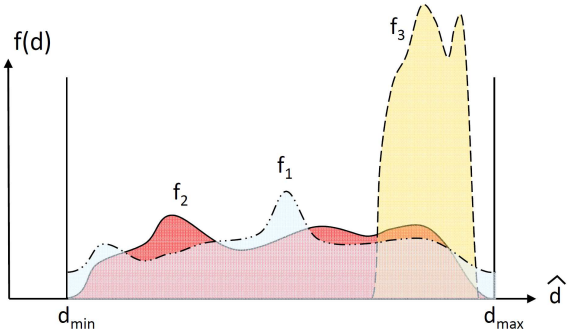

Figure 2 - Different depth-range settings for the same scene. f_1: overly-narrow-tuned depth range, f_2: optimally tuned depth range for high resolution, f_3: overly-wide-tuned range

It is evident that a properly tuning of the depth interval all the time to get a function with properties like f_2 must be the goal for range setting. However in so doing, the scene could change so rapid that a function like f_2 may turn also fast into a function like f_1, which is undesirable.

Therefore the first step of the approach should be to modify the range extremes for detecting in advance crosses of the probability mass through them. Thus, two new limits for the extremes are defined. They are referred to as soft limits $\delta_{\min}$ and $\delta_{\max}$ in Fig. 3.

The soft limits are set up regarding the discretization of the inverse depth range as $\delta_{\min} = d_{\min} + m_{\min}\Delta$ and $\delta_{\min} = d_{\max} - m_{\max}\Delta$, where Δ is a discretization unit of the inverse depth interval, $m_{\min}$ and $m_{\max}$ are positive integers, for instance $m_{\min} = m_{\max} = 3$ for $q = 32$, it is about 10% of q.

On one side, the extent of the mass flowing outwards suggests the reaction speed to modify the range span since the camera begin to focus parts of the scene wherein objects are appearing ahead and/or behind the cost volume boundaries. On the other side the amount of the mass outside the soft limits suggests the magnitude of the correction.

Finally, for a fast reaction of the approach, it is more pressing the case in where the largest portion of mass approaches fast to the extreme $d_{\max}$ with maximal velocity than any other case. Thereby, the scattering on mass along with its velocity field have to be cautiously combined in order to define the strength of correction effectively.

B. Adaptive correction

In order to compound the mass rate and amount, two elements are introduced to conform an approach. The first one are the weighting functions $W_{\min}$ and $W_{\max}$ on the soft intervals $[d_{\min}, \delta_{\min}]$ and $[\delta_{\max}, d_{\max}]$, respectively. They penalize the probability mass on the soft intervals. These are

$$W_{\min}(d_{\min} + i\Delta) = K_{\min} e^{-n(\delta_{\min} - d_{\min})i\Delta} \quad \text{for } i = 0, 1, \cdots, \delta_{\min} \tag{3}$$

$$W_{\max}(d_{\max} - i\Delta) = K_{\max} e^{-n(d_{\max} - \delta_{\max})i\Delta} \quad \text{for } i = 0, 1, \cdots, \delta_{\max} \tag{4}$$

where $K_{\min}$ and $K_{\max}$ are gain factors for the penalization, n is a non-negative integer to strength the exponential decay, $(\delta_{\min} - d_{\min})$ and $(d_{\max} - \delta_{\max})$ are coefficients of the exponential functions.

Next, we can define flow fields for the mass escaping from inside out of $[\delta_{\min}, \delta_{\max}]$. The flow fields are calculated by the transversal mass velocity on $[d_{\min}, \delta_{\min}]$ and $[\delta_{\max}, d_{\max}]$ as

$$v_{\min}(d_{\min}+i\Delta) = f(d_{\min}+i\Delta)\text{-}f(d_{\min}+(i+1)\Delta) \quad \text{for } i=0, 1, \cdots, \delta_{\min} \tag{5}$$

$$v_{\max}(d_{\max}\text{-}i\Delta) = f(d_{\max}\text{-}i\Delta)\text{-}f(d_{\max}\text{-}(i+1)\Delta) \quad \text{for } i=0, 1, \cdots, \delta_{\max}. \tag{6}$$

It is noticing that every rate element of the field may be positive or negative according to the direction of the elemental mass flow wherein the sign is positive when the mass element flows outwards and vice-versa. Finally, we sum up the weighted flow field at both sides of $[\delta_{\min}, \delta_{\max}]$ as

$$F_{\min} = \sum_{i=0}^{\delta_{\min}} W_{\min}(d_{\min} + i\Delta)\ v_{\min}(d_{\min} + i\Delta) \tag{7}$$

$$F_{\max} = \sum_{i=0}^{\delta_{\max}} W_{\max}(i\Delta + \delta_{\max})\ v_{\max}(i\Delta + \delta_{\max}). \tag{8}$$

Now we are in the position to establish a procedure for calculating the range modifications in real time. The approach consists in modify $d_{\min}$ and $d_{\max}$ and recalculate the new range and Δ. We will set down the algorithm for $d_{\min}$ solely, because the counterpart for $d_{\max}$ is similarly deducible.

The approach for correcting the lower part of the range consists of two differentiated parts. The first one includes the case when the probability mass is scattered on the soft interval while the second part deals with the case

when the probability mass is on a narrow subinterval. Correspondingly the first part of the algorithm is

$$\begin{aligned}
&\text{IF } f(d_{\min}+i\Delta)>0 \text{ on } [d_{\min},\delta_{\min}] \text{ THEN} \\
&\quad \text{Calculate } W_{\min}(d_{\min}+i\Delta) \\
&\quad \text{Calculate } v_{\min}(d_{\min}+i\Delta) \\
&\quad \text{Calculate } F_{\min} \\
&\quad \eta = f_r(c_{\min}F_{\min}/\Delta) \text{ and } d_{\min} := d_{\min}-\eta\Delta \\
&\quad \text{IF } d_{\min}<0, \text{ THEN } d_{\min}:=0 \text{ ENDIF} \\
&\quad \Delta := (d_{\max}-d_{\min})/q \\
&\text{ENDIF}
\end{aligned} \tag{9}$$

where $f_r(.)$ is the floor function, η is a nonnegative entire number, $c_{\min}$ is a gain constant that adjust the correction strength. The second part of the approach accomplishes the following operations

$$\begin{aligned}
&\text{IF } F_{\min}=0 \text{ THEN calculate } [\delta_{\min}+m\Delta] \\
&\qquad \text{wherein } f(\widehat{d}) \text{ is identically zero} \\
&\quad \text{IF } m>0 \text{ THEN} \\
&\qquad\quad d_{\min} := d_{\min}-m\Delta \\
&\qquad\quad \Delta := (d_{\max}-d_{\min})/q \\
&\text{ENDIF ENDIF}
\end{aligned} \tag{10}$$

VII. Case Studies

The proposed approach is tested in real scenarios and datasets pertaining to the Computer Vision Group at the TU Munich. We summarize the results for a dataset with an IR depth-finding camera Kinect sensor. The kinect sensor employes color gradients from white (near) to blue (far) to depict the depth map. Similarly, in our map estimation a gray-scale gradient from white (near) to black (far) is used to visualize depth.

Fig. 3 shows the outcomes of map estimation indoors with he adaptive approach. The raw frame complements with the depth map of the kinect sensor for a 3D composition of the world in real scale. One ascertains that the histogram spans the whole inverse depth range with decay for very short distances (right limit) and is tuned for infinitely distant points (left limit). In broad outlines, the estimation is very truthful up to scale in comparison with the accurate depth measures of the kinect sensor. However, one notices both in the regularized map and the only-data-term-based map the presence of some white stains in the background. This fact indicates a great depth uncertainty for far physical points. This drawback was observed in many applications of DTAM which warns of a large sensibility to small parallax at least in the way that DTAM is commonly adopted.

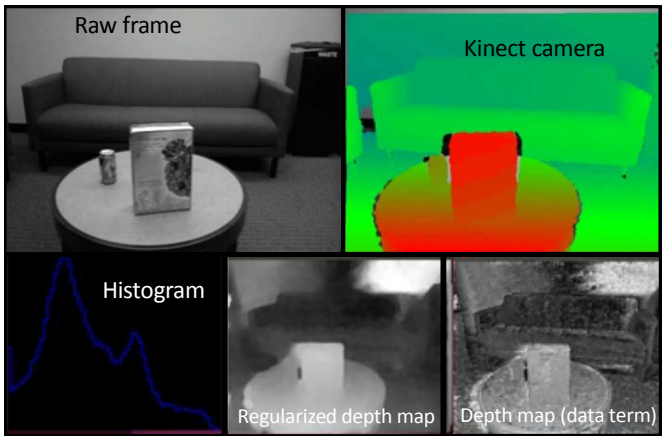

Figure 3 - Map estimation with adaptation of the inverse depth range. Comparison with kinect camera sensor

Wide differences in results exist comparatively between automatically adapted and fixed depth ranges. For instance in Fig. 4, the histogram is set up too tight for the scene. So the right extreme of the histogram begins to increase when more and more physical points approach to the camera. This, clearly will denote false estimations at short distances. Thereby they are countless white specks and great white spots observed in the depth map with data term. The regularized map illustrates the same drawback as well, although even more serious, since edges and contours yield blurred.

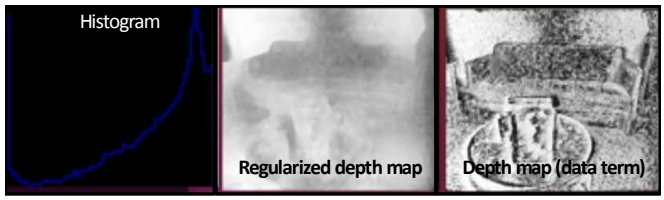

Figure 4 - Map estimation with fixed narrow inverse depth range

Fig. 5 depicts the case of fixed, but on the contrary as former case, wide range. The depth map states a significant improve with respect to the previous case. Even when the tuning is appropriate for the scene, it seems from its histogram that it is somewhat conservative for short distances (see the strong decay of the histogram a bit far from the right extreme). This result in a loss of some details of the close by scene.

Fig. 6 portrays the evolution of the adapted depth range, showing a permanently adjustment of the upper limit, it is short distances. This adaptation proves to be more effective in comparison with the fixed set up of a wide range.

The most categorical indicator of the estimation quality is given in Fig. 7 for the mean errors of the estimated depths. Clearly the adaptation of the depth range yields the lowest error.

Figure 5 - Map estimation with fixed wide inverse depth range

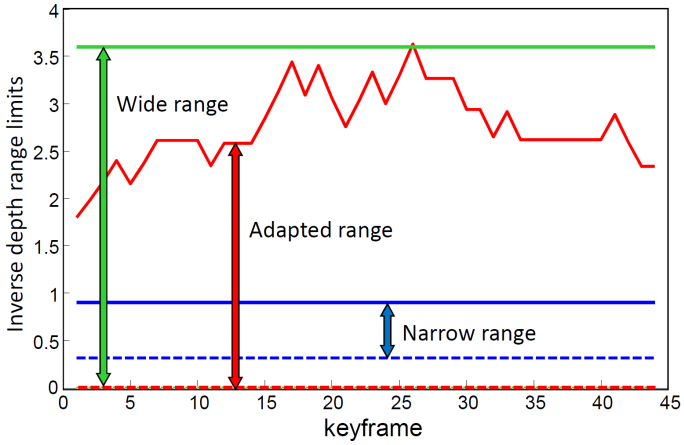

Figure 6 - Time evolution of inverse depth range limits

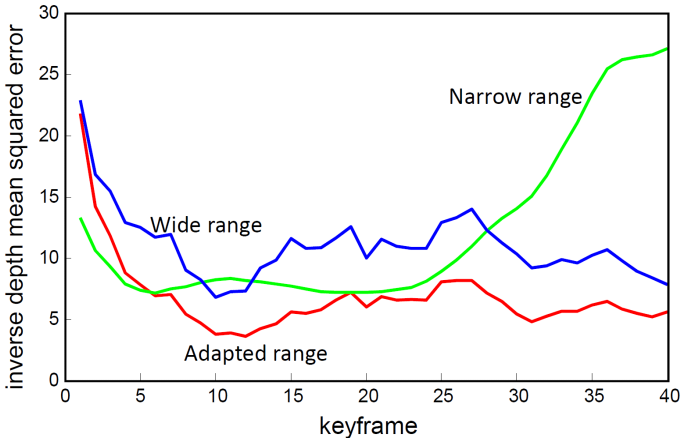

Figure 7 - Quality of map estimation with fixed and adapted depth range

VIII. Conclusions

In this work we have presented an original approach to adapt the depth range automatically for world-scale changing environments. The approach focuses first on the problem of the inherent discretization error of the depth range in the cost volume and second on the estimation errors that can befall when objects rests outside of the set depth range. Consistently it sheds light in the mechanism whereby the estimator manages the photometric consistency and from this issue it elaborates an approach to optimally adapt the depth range. This is based depth histogram of the keyframe and the distorting of the probability mass in time. The adaptation is based in restraining the mass to spread out of the range interval by expanding or shrinking the range span to contain the mass inside.

This outcomes impacts positively in the performance of SLAM namely in the accuracy of the camera tracking. Many other examples with datasets in air and underwater (which were not shown here for the sake of being succinct), provide solid evidence of the high robustness which is achievable in harsh scenarios, described for instance by characteristics of self-similarity, blurriness and low-level lighting.

References

[1] S. Lucey. *Direct Visual SLAM*. In 16-623 - Designing Computer Vision App. Carnegie Mellon. The Robotics Institute, 2016.

[2] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. Journal of Mathematical Imaging and Vision, 40(1):120–145, 2011.

[3] R.A. Newcombe, S.J. Lovegrove, S.J. and A.J. Davison. DTAM: Dense Tracking and Mapping in Real-Time. In *2011 IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2320-2327. IEEE, 2011.

[4] J. Engel, J. Schöps, and D. Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *European Confonference on Computer Vision (ECCV)*, pages 834-849. Springer, 2014.

[5] R.A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A.J. Davison, P- Kohi, J. Shotton, S. Hodges and A. Fitzgibbon. Kinectfusion: Realtime Dense Surface Mapping and Tracking. In *10th IEEE Int. Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 127-136. IEEE, 2011.