Robust Pose Estimation Based on Normalized Information Distance

Zhaozhong Chen and Christoffer Heckman*

***Abstract*— Dense image alignment works by minimizing the photometric error of two images since it is assumed that the illumination changes between images close in time remain the same—this is what is called the brightness constancy assumption. However, this assumption does not hold with long-term maps since illumination changes continually from day to day (morning, afternoon, evening) and is dependent on certain external conditions like weather or even seasons. In this work, we present an image registration algorithm based on the Normalized Information Distance (NID) that is shown to be robust to extreme illumination changes comparing to the traditional direct methods. The pose is estimated by minimizing the NID function with the help of the nonlinear least square optimization library G2O. We share our source code[1] (CPU and GPU version) for the benefit of the community, which can be a strong basis for future tracking and mapping system based on NID.**

I. INTRODUCTION

Direct visual odometry [1], [2] works by minimizing the photometric error of two images since it is assumed that the illumination changes between images close in time is small—this is what is called the *brightness constancy assumption*. This works well for visual odometry, which estimates the 3D pose of the camera frame-to-frame, since the changes in the scene's lighting is minimal within a short time interval, especially with a high frame rate camera.

Frame-to-frame tracking, however is not ideal due to natural drift over time. Instead, it is desirable to localize the camera with respect to a prior map. This is the essence of simultaneous localization and mapping (SLAM): not only to map an environment but also to localize against it. However, the brightness constancy assumption does not hold with outdoor long-term maps since illumination changes continually from day to night, it doesn't hold with indoor illumination change such as light source turn on and turn off either. As such, a direct photometric minimization is often inadequate in these kind of situations. A technique that is capable of registering against a map and is robust to illumination changes like those shown in Figure 1 is of great interest to roboticists. In this paper, we propose a dense image alignment algorithm based on information entropy that is robust to extreme illumination changes while maintaining the benefits of typical direct photometric image alignment methods.

Both authors are with the Autonomous Robotics and Perception Group, at the University of Colorado at Boulder. ⋆ christoffer.heckman@colorado.edu

[1] https://github.com/arpg/NID-Pose-Estimation

Fig. 1. Example images of the similar scene with different illumination condition, globally and locally. (ETHZ CVG illimunation change dataset).

II. RELATED WORK

While some work exists to create maps that are capable of coping with extreme illumination changes, like for example experienced based mapping [3] or localizing against a navigation sequence [4], these techniques mitigate the problem by simply using redundant information and storing all the different conditions in maps that are co-registered. These approaches are not suitable for real world implementation.

By far, the most common techniques involve folding in the illumination robustness directly into the image registration optimization. An example of such a technique is applying Zero Normalized Cross Correlation (ZNCC) [5], [6] by which the scene's mean intensity is subtracted to each pixel which in turn is then divided by the standard deviation. This aids in adjusting the brightness of the image due to variations in lighting and exposure conditions. Another similar technique is the Global Affine Illumination (AI) [2]. In this method, two extra parameters are added to the tracking optimization: a scale α and a bias β parameter that is applied to each pixel's intensity when calculating the photometric error.

These techniques have a major drawback in that they assume that the illumination change is globally consistent. It is often the case that lighting conditions affect the scene differently in different areas, depending on the geometry of the scene (e.g. multi-path), the camera position (e.g. non-Lambertian reflections [7]) or the material properties of the objects within it (e.g. albedo).

One error function that is robust against this non-uniform lighting limitation, and is one of the most popular techniques used in computer vision, is called the *Census transform* [8],

[9]. Whereas the Sum of Absolute Difference (SAD), or even the Sum of Squared Differences (SSD) directly operate over the photometric values of pixels, Census converts each pixel into a binary signature that encodes whether a pixel's photometric value is lower or not compared to its neighboring pixels. This has the advantage that it encapsulates local consistencies of illumination changes, rather than assuming a global illumination transform. However, given the fact that each pixel is now a binary signature, a direct subtraction can no longer be applied to find correlations. Thus, the Hamming distance is used as a metric of similarity by which the number of matching bits in the binary signature are counted in order to provide the final score. A major disadvantage of using the Hamming distance in optimizations is that it is not continuous. Thus, what the authors of the work in [10] proposed is to use each bit independently as a different channel or *plane*. Furthermore, each bit is interpolated during the reprojection in order to find the error per bit which is then aggregated with the rest of the Census bits to calculate the final error for that pixel. The downside to using this technique, however, is that the computation per pixel has now been incremented by the number of bits in the Census transform used. Other work such as [11] novely trains a network to recover the image to a canonical appearence from different illumination condition. However, this approach is limited by its training dataset.

A number of recent works, e.g. [12] in the real-time frame-to-frame tracking problem have focused on *indirect* methods, i.e. those that compute features over the image in order to conduct alignment. These methods are highly popular for many applications of frame-to-frame tracking in particular, e.g. dense map fusion [13]. Furthermore, features may be designed to be considerably illumination invariant [14], and with deep learning techniques, can be run in real-time on an available GPU. However, in this work we focus exclusively on the direct problem in order to explore the potential upside in this particular domain.

More recently, the Normalized Information Distance (NID) metric [15], [16] was proposed. This technique makes use of mutual information [17], [18], [19] which has been shown to be robust in the alignment of multi-modal images, in particular for medical applications. The images are converted into histograms, which are then aligned using an entropy metric. This method, although far superior to others, is extremely computationally intensive and can take up to three orders of magnitude more than the typical photometric minimizations; even those with robust lighting techniques. To increase the efficiency NID for pose estimation, [20], [21], [22], [23] investigate using second order optimization method to minimize the cost function. It is extremely useful in other SLAM algorithms like pose graph relaxations [24], [25] or sensor fusion [26]. [23], [22] use Mutual Information instead of NID as the metric. However, the mutual information is not a true metric. This problem is further discussed in Section III-B. [20], [21] use NID as well as second order optimization algorithm to accomplish tracking task and achieve impressive result. However, they use the whole image to generate one residual for optimization, which make the traditional second order optimization strategy such as Levenberg-Marquardt(LM) being suboptimal. LM is more efficient when the residual number is higher than the number of parameters. In their case, they only have one residual but have 6D pose to estimate. Comparing to those previous work, our main contribution is stated as follows.

- We firstly investigate grid cell approach for NID calculation using LM method.
- While B-spline functions have been used in previous approaches to NID, [20], [21] we are the first to investigate their parameters' influence on the optimization result.
- We derive the detail formula for the NID method and release our code under a standard LM optimization framework using G2O.

To the best of our knowledge, there is no open source code using NID for 6D pose optimization. In the remainder of this paper, Section III introduces the definition of NID in information theory. Section IV list our cost function, demonstrate the benefit of the cell based approach, and derive the Jacobian detail for the LM opimization. Section V compares the NID method with the tradictional direct method.

III. BACKGROUND

A. Mutual Information

The concept of entropy and mutual information [27], [28] have already been widely discussed in information theory. Entropy H of a random variable X is defined as $H(X) = \sum_{x \in X} P_X(x) I_X(x)$, where $P_X(x)$ is the probability mass function and $I_X(x) = -\log\left(P_X(x)\right)$ is the information content. Entropy can also be extended to two random variables to measure the amount of information obtained about one random variable, through the other random variable. This is what is called mutual information. Figure 2 shows a graphical representation of two random variables and their entropy. $H(X)$ and $H(Y)$ are the individual entropies for the correlated random variables X and Y. $H(X,Y)$ is the joint entropy while $H(X|Y)$ and $H(Y|X)$ are the conditional entropies. The purple is the mutual information $I(X,Y)$. Formally it is defined as:

$$I(X,Y) = H(X) + H(Y) - H(X,Y), \tag{1}$$

where $H(X,Y) = -\sum P_{XY}(x,y)\ \log\left(P_{XY}(x,y)\right)\ \forall\ x \in X,\ y \in Y$. As can be seen from this definition, mutual information is maximized as the two circles overlap completely. In contrast, variables that are independent have circles that do not overlap and as such convey no mutual information. For image X, the probability of a pixel x_i being of a certain value $I(p)$ is the number of pixels at that value divided by the total pixels M, i.e.:

$$\begin{aligned} P(x_i) &= \frac{1}{M} \sum \beta(I(p)) \\ \beta(I(p)) &= \begin{cases} 1 & I(p) = x_i \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

B. Normalized Information Distance (NID)

One of the biggest issues with mutual information is that it is not a true metric as it does not satisfy the triangle inequality. Consider again the case of Figure 2; in this example, both cases have the same amount of mutual information, yet with different joint entropies. It is desirable to disambiguate between the two cases, and have a distance that captures $\mathcal{D}(X_1, Y_1) > \mathcal{D}(X_2, Y_2)$.

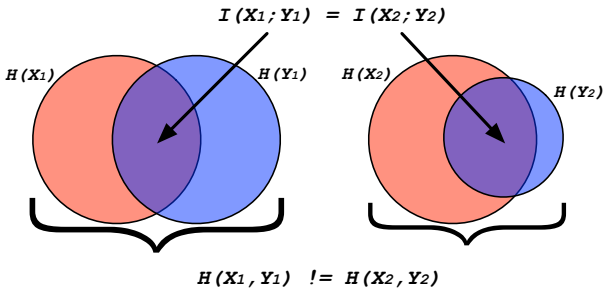

Fig. 2. Entropy diagrams of two non independent random variables. The mutual information is denoted as $I(X;Y)$. In these two cases, the mutual information is the same $I(X_1;Y_1) = I(X_2;Y_2)$ (purple areas), yet the joint entropies are different $H(X_1, Y_1) \neq H(X_2, Y_2)$ (captured by the brackets).

The NID is designed to overcome this limitation by normalizing the variation of information by the joint entropy; furthermore, unlike mutual information, it is a true metric. Formally, NID is defined as:

$$NID(X,Y) = \frac{H(X,Y) - I(X,Y)}{H(X,Y)}. \tag{3}$$

IV. METHOD

Pascoe et al. [20], [29] were one of the first to propose a dense tracking system that directly minimizes a single global NID cost. The optimization uses the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, which is a quasi-newton method by which the Hessian is iteratively approximated from the gradient and provides robustness by using a line search. In their method, the images are converted into a single histogram, inconsistencies due to outliers cannot be handled like typical direct photometric methods could. This work proposes to use NID as an image distance measure instead that provides multiple residuals, and therefore the problem can be rewritten to use the conventional Lucas-Kanade [30] whole image alignment formulation. As such, a second-order optimization method like Levenberg–Marquardt (LM) [31] can be used. Furthermore, the optimization is now capable of providing its true covariance matrix. To achieve this, the image is now split into cells, for each of which a NID cost is computed. The optimization is looking to minimize the *summation* of the NID costs for all cells in the image. The cost function of whole image alignment is written as follows:

$$\arg\min_{\mathbf{T}_{cur}} \mathcal{D}\left(I_{cur}, \omega^{-1}(\mathbf{T}_{cur}^{-1}, \omega(I_{ref}, \mathbf{T}_{ref}))\right) \tag{4}$$

where $\mathbf{T}_{cur} \in \mathbb{SE}(3)$ is the six degree-of-freedom pose to be estimated from camera frame to world frame. $I_{cur} : \Omega \to \mathbb{R}^+$ and $I_{ref} : \Omega \to \mathbb{R}^+$ are the current and reference images respectively, and ω is the warping function that transforms the reference image using the known parameters $\mathbf{T}_{ref}$ and camera intrinsics $\mathbf{K}$ to the 3D world frame. Then we use the estimated $\mathbf{T}_{cur}^{-1}$ and warping function to project 3D world points back to the current image frame. The initial guess of $\mathbf{T}_{cur}$ can be infered from the constant velocity model as we know the $\mathbf{T}_{ref}$ and its previous frame's transformation from local to world.

Finally, $\mathcal{D}$ in Eq. (4) is a measure of distance and as such must be designed to meet certain conditions:

- Non-negativity: $\mathcal{D}(X,Y) \geq 0$
- Equivalence: $\mathcal{D}(X,Y) = 0 \iff X = Y$
- Symmetry: $\mathcal{D}(X,Y) = \mathcal{D}(Y,X)$
- Triangle Inequality: $\mathcal{D}(X,Y) + \mathcal{D}(Y,Z) \geq \mathcal{D}(X,Z)$

This work posits that the NID operation defined in Eq. (3) satisfy the above requirement and provides a robust result to brightness inconstancy so we use Eq. (3) as $\mathcal{D}$: $\mathcal{D}(X,Y) = NID(X,Y)$

Intuitively, $H(X)$ and $H(Y)$ are the individual entropies of the images X and Y whose distributions are calculated by Eq. (2). However, what we practically use to calculate the distribution is a B-spline function, which overcomes the limitation that Eq. (2) is not differentiable. Details on this are discussed in the next subsection. The NID cost only requires computing the joint distribution, since the individual distributions can be calculated by marginalization. A sampling method is performed to construct each distribution, represented as a N-bin (interval) histogram, with the final joint distribution being of size $N \times N$.

Finally, the spatial distribution of cells in the image

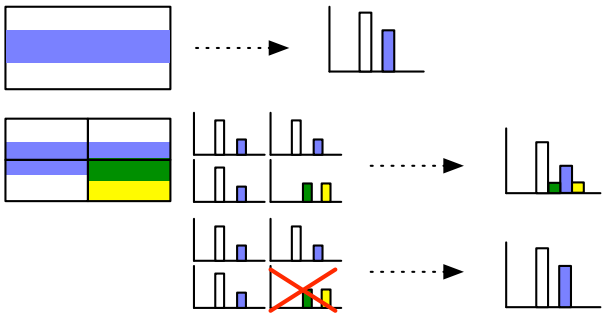

Fig. 3. The grid cell approach preserves the histogram consistency regardless of reprojection errors due to incorrect depth estimates or outliers. For the case of the whole image NID, the final histograms between the top and middle row will not match correctly. With the grid cells, however, the cells that are affected can easily be down-weighted or rejected by the huber norm in least square optimization and as such preserve the original histogram (bottom row).

provides additional benefits with regards to histogram consistency. For the single residual NID case, the joint entropy will possibly be the minimum at some incorrect transform since the bad observations were already added to the histogram. However, with the cell based approach the position of the black area of the image will only affect the cells it reprojects into, and as such, the histograms of the other cells will not be affected (Figure 3).

A. Optimization Strategy

We use the G2O [32] as the backend to implement the LM algorithm. Two terms need to be defined: the first is the error term, which is described as Eq. (4); and the second is the Jacobian matrix, i.e. the Jacobian of NID w.r.t. $\mathbf{T}_{cur}$. For brevity, we rewrite $\mathcal{D}\left(I_{cur}, \omega^{-1}(\mathbf{T}_{cur}^{-1}, \omega(I_{ref}, \mathbf{T}_{ref}))\right)$ from Eq. (4) as $D(I_c, I_r)$. To provide clarity frequently absent from previous works, we explicitly derive the Jacobian:

$$\frac{\partial(D(I_c, I_r))}{\partial \mathbf{T}_{cur}} = \frac{\frac{\partial H(I_c, I_r)}{\partial \mathbf{T}_{cur}}(H(I_c) + H(I_r)) - \frac{\partial H(I_r)}{\partial \mathbf{T}_{cur}} H(I_c, I_r)}{H^2(I_c, I_r)}. \tag{5}$$

For brevity, we rewrite $H(I_c, I_r)$ and $H(I_r)$ as $H(\cdot)$. According to the chain rule, $\frac{\partial H(\cdot)}{\partial \mathbf{T}_{cur}}$ can be writen as:

$$\frac{\partial H(\cdot)}{\partial \mathbf{T}_{cur}} = \frac{1}{M} \sum \frac{\partial H(\cdot)}{\partial P(\cdot)} \frac{\partial P(\cdot)}{\partial \beta(\cdot)} \frac{\partial \beta(\cdot)}{\partial I(\cdot)} \frac{\partial I(\cdot)}{\partial p} \frac{\partial p}{\partial \mathbf{T}_{cur}} \tag{6}$$

For full detail on $\frac{\partial H(\cdot)}{\partial P(\cdot)} \frac{\partial P(\cdot)}{\partial \beta(\cdot)}$, see [20]. Note that $\frac{\partial I(p)}{\partial p}$ is the pixel intensity gradient and the $\frac{\partial p}{\partial \mathbf{T}_{cur}}$ is the projected pixel's derivative w.r.t. the $\mathbb{SE}(3)$ pose. Generally we need to transform it to calculate the derivatives on the perturbation of its lie algebra $\mathfrak{se}(3)$. The detail is in [33], [34] and for simplification it is given by:

$$\frac{\partial p}{\partial \mathbf{T}_{cur}} = \begin{bmatrix} \frac{-f_u x y}{z^2} & f_u + \frac{f_u x^2}{z^2} & -\frac{f_u y}{z} & \frac{f_u}{z} & 0 & -\frac{f_u x}{z^2} \\ -f_v - \frac{f_v y^2}{z^2} & \frac{f_v x y}{z^2} & \frac{f_v x}{z} & 0 & \frac{f_v}{z} & -\frac{f_v y}{z^2} \end{bmatrix},$$

where f_u, f_v are the camera focal length, and with a slight abuse of notation, x, y, z represent the 3D point location in current camera frame after being projected from the world frame by $\mathbf{T}_{cur}$. Note that z, the projective depth of a pixel, may be either measured directly (e.g. in RGB-D) or conducted with arbitrary scale through inference (e.g. [35]). However, $\beta(\cdot)$ from Eq. (2) is not differentiable. Hence, we use a basis of B-splines to take its place while maintaining $\beta(\cdot)$ function's constraints. B-splines have basic spline function properties deriving from their piecewise polynomial origins. However, they also have unique behaviors that fit our purposes, which will be detailed below. B-splines are calculated recursively and can be written as [36], [37], [38]:

$$\beta_{i,k}(t) = \begin{cases} 1 & t \in [t_i, t_{i+1}) \\ 0 & \text{otherwise,} \end{cases}$$

if $k = 1$ and, if $k > 1$,

$$\beta_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} \beta_{i,k-1}(t) + \frac{t_{i+1} - t}{t_{i+k} - t_{i+1}} \beta_{i+1,k-1}(t),$$

where t_i is B-spline knots, k is the B-spline degree. For $k > 1$ the B-spline function is differentiable because it is a k-degree polynomial and its derivative is:

$$\frac{d\beta_{i,k}(t)}{dt} = \frac{t - t_i}{t_{i+k-1} - t_i} \frac{d\beta_{i,k-1}(t)}{dt} + \frac{t_{i+1} - t}{t_{i+k} - t_{i+1}} \frac{d\beta_{i+1,k-1}(t)}{dt} + \frac{1}{t_{i+k-1} - t_i} \beta_{i,k-1}(t) - \frac{1}{t_{i+k} - t_{i+1}} \beta_{i+1,k-1}(t) \tag{7}$$

The knots of a B-spline in our application are arranged according to the bin number N and B-spline degree k. The degree k is fixed to 4, a choice discussed in Section V-B, and the influence of different bin numbers N is also shown and discussed later. After defining the bin number and B-spline order, we can obtain the knot vector according to:

$$t_i := \begin{cases} 0 & if \quad i < k \\ i - k + 1 & if \quad i \leq k \leq N \\ N - k + 1 & if \quad N < i < N + k \end{cases} \tag{8}$$

The arrangement of the B-spline function as described in Eq. (8) is known as the clamped knot vector. Note that this has some duplicate knots such as the 0 in the Eq. (8). With the help of clamped knot vector, we maintain by construction $\sum_{i=0}^{N+k-1} \beta_{i,k}(t) = 1$. This is important because after we sum up the B-spline values of all the pixels in a cell and divided by the pixel number M, the uniform value of the probability mass function summation still holds.

Finally, as the knot vector range is between $[0, N-k+1]$, we need to map the pixel intensity into knot vector range to compute its B spline value, by:

$$t = I(p) \frac{N - k + 1}{255}. \tag{9}$$

Combining Eq. (7) and Eq. (9) we can have the $\frac{\partial \beta(\cdot)}{\partial I(\cdot)}$ in Eq. (6).

V. EXPERIMENTAL RESULTS

A. Evaluation Method

To evaluate the algorithm, experiments were conducted on the synthetic ETHZ-CVG illumination change dataset [39] and TUM RGB-D dataset [40]. [39] provides image pixel intensity varies globally and locally (like Figure 1).

We want to demonstrate that the proposed method can give a more accurate pose estimation compared to a traditional direct method when the lighting condition is not consistent. We conduct three different traditional direct methods experiment for comparison. We perform our simple direct method by minimizing the photometric error of an image's high gradient pixels [41], [35], i.e.:

$$\arg\min_{\mathbf{T}_{cur}} ||I_{cur} - \omega^{-1}(\mathbf{T}_{cur}^{-1}, \omega(I_{ref}, \mathbf{T}_{ref})||^2. \tag{10}$$

We then run the same dataset on the semi-direct library SVO [22] and direct method library DSO [35]. We must note that our current library focus on estimating the 6D pose between an image pair instead of the whole image sequence so we cannot run our algorithm on the whole dataset and compare the trajectory error with DSO or SVO. Instead, we only use

the LM algorithm part of DSO and SVO to estimate the poses between image pairs like our NID algorithm. To make the comparison fair, we don't allow depth updation in DSO and SVO. Instead, we feed them with known depth from the RGB-D dataset and fix the depth. We focus our work on the comparison of different conventional direct methods to demonstrate its imporvements on different illumination conditions.

For the NID method, we split the image into different cells, since the number of cells into which the image is split will affect the final result. Table I shows the number of splits for each configuration, and the corresponding number of cells and cell size. One split means the image is divided by two in each dimension, yielding four cells, and so on.

In the global illumination condition varying subset, the whole image sequence goes from bright to dark and dark to bright continuously over multiple frames, simulating turning a light off and on. In the local illumination condition varying subset, only a sub-part of the image's illumination changes. Of thoses datasets, we choose 90 pairs image each, with each pair having a different illumination condition like Figure 1. For example, in subdataset *ethl_syn1_global*, we choose image pair {0035.png, 0040.png}, {0036.png, 0041.png}... Such a pair has a strong illumination difference directly from visualization. For each image pair, we choose one camera pose as static reference $\mathbf{T}_{ref}$ and the other one is the to be optimized pose $\mathbf{T}_{cur}$. Both poses are from the groundtruth. Then we add a fixed disturbance in rotation and translation $\mathbf{T}_{noise}$ to the $\mathbf{T}_{cur}$ for all the image pairs to imitate the none-perfect pose estimation from a tracking system.

TABLE I
NID CONFIGURATIONS

Configuration	Splits	Total Num Cells	Resolution of Cell
NID1	1	4	320x240
NID2	2	16	160x120
NID3	3	64	80x60
NID4	4	256	40x30
NID5	5	1024	20x15

Finally we apply NID method and the direct method to optimize the $\mathbf{T}_{cur}$ with noise seperately. In both datasets, the noise we add to the groundtruth is 0.03 meters in translation and 0.015 radians in rotation because the average translation and rotation movement between two consecutive images are around this two numerical values. The way we calculate the rotation error is by mapping the rotation from Lie group $\mathbb{SO}(3)$ to the Lie algebra $\mathfrak{so}(3)$ so that we can have a minimal representation of the 6D estimated pose and ground truth, then we calculate the root sum-of-squares of the difference between the estimated pose and ground truth. To further investigate the numerical result, we define "successful estimation" if the optimized pose's root sum-of-squares rotation and translation error from ground truth is smaller than $\delta = 0.045$. This is not a strict metric, however, if the optimized error is smaller than δ, it generally means the difference between the noise pose and groundtruth gets smaller after optimization. It should be noted that NID1 and NID2 are not tested since they would result in too few residuals and not suitable for LM algorithm. Also, The cell resolution naturally is reduced to the point that no clean split can be performed after 5 splits. Thus, experiments were performed between 3 to 5 splits, yielding 64 to 1024 residuals.

B. Results

1) ETHZ CVG dataset test: The left and middle side of Figures 5 show the median/max/min translation and rotation errors of the 90 ETHZ-CVG dataset image pairs with global illumination change after optimization, the midlle figure shows the same error statistics but using the local illumination change dataset. We can see the NID method here shows its advantage mostly in improved translation results. The translation of the direct method has a large varaince while NID4 and NID5 is quite robust. Interestingly, the direct method's rotation estimation is stable but the NID4 still surpasses it with smaller median/max/min error.

Table II shows the percent of successful estimates of the direct method and NID method with different number of cell splits as defined in the previous section, where the pose estimation error falls below a fixed threshold. "G" stands for the global illumination varying dataset and "L" represents the local illumination varying dataset. Using the result from "G" row, when there is a significant illumination change between the images, the direct method fails at most of the image pairs with a only 12.22% successful rate. With NID, the success rate increases significantly, with NID4 returning the higest success rates. When the cell number is too small, for example the NID3 can only yield 64 measurements, the system is not found to be robust to the outliers of larger NID values. However, when the cell number is too large like NID5, a cell's pixel number (300) is small, which means it may not be able to provide valid information over a wide enough receptive field. In both "G" and "L", NID4 shows the highest succesful rate.

TABLE II
SUCCESSFUL ESTIMATES FOR ETHZ-CVG DATASET

	DIR	DSO	SVO	NID3	NID4	NID5
G	12.22%	22.22%	14.45%	34.44%	**80.00%**	53.33%
L	13.30%	45.56%	22.21%	53.33%	**74.44%**	48.89%

2) B spline coefficients: There are several customized variables that may influence the errors; for instance, setting the B-spline order to $k = 4$ ensures the cost function has a C^2 continuous derivative. We also show results for the number of bins, a frequently-unchanged setting which we found has impact on the results of the estimation. Following Eq. (8), we let $k = 4$ and $N = 6, 8, 10, 12, 14$ respectively, then run the NID method with 90 image pairs same as previous subsection V-A. The split number is fixed to 4 in different bin numbers comparison.

Table III shows the percent of successful estimates (as defined in the previous section); the right plots in Figures

TABLE III

SUCCESSFUL ESTIMATES WITH DIFFERENT BIN NUMBERS

6	8	10	12	14
67.78%	80.00%	**81.11%**	65.56%	72.22%

5 shows the translation and rotation error box plot for the different bin numbers. We can see that, unlike the image cell number setting, the bin number doesn't affect error very significantly. The translation and rotation median/max/min error are comparable, although $N = 10$ yields the best result. When we use NID4, a small bin number (extreme example being $N = 1$) or a large bin number (extreme example being $N = 255$) may both lead to data losing significant information content. Therefore we design the number of bins N to have each cell's pixel count set such that each cell can have a resonable number of pixels, which is dependent on the resolution of the image and therefore reduces to a choice based on application. In the case of NID4, our tests show $N = 8$ or 10 can yield the smallest pose estimation error.

3) TUM RGB-D dataset test: The TUM RGB-D dataset doesn't have illumination change. We simply imitate the ETHZ-CVG dataset and artificially add illumination change to the dataset. We continuously add and subtract intensity to each pixel. Figure 4 shows a sample of the fr1/desk image sequence of the TUM RGB-D dataset after adding illumination change. Again, we choose 90 image pairs from different TUM RGB-D datasets, add the same noise to the groundtruth pose as the ETHZ-CVG dataset, and run the pose optimization algorithm from different methods. Due to the page limitation, we only show the succesful rate in Table IV. It's not a surprise that the traditional methods have poor performance on the real world dataset with rolling shutter camera. For example, the SVO successful rate is 0 on every dataset so we don't list it on the table. Although we didn't compare with the DSO and SVO on the whole image sequence, we tried to run them to visualize the final trajectory. SVO cannot initialize on any of the illumination change TUM and DSO is struggling at initialization and can barely initialize successfully. This behavior corresponding to the table result implicitly.

Fig. 4. We manually change the illumination condition of the TUM RGB-D dataset. We firstly decrease the light intensity, then increase it and do the process continuously.

VI. CONCLUSION

In this work, we introduced an algorithm that makes use of the Normalized Information Distance (NID) metric for whole image alignment. The algorithm splits the image into cells, each counting as an observation for a least squares style LM optimization. The NID metric was shown to be robust to strong illumination changes compared to conventional photometric algorithms. Also, we are the first to open source vision-based NID localization code for pose estimation. The CPU version can take 2 to 3 min to optimize the pose while the GPU version can be 100 to 300 times faster.

In all, this work is a basis for a future open-source NID-based tracking and mapping system. There are multiple avenues for future work resulting from this effort. First, to the best of our knowledge, sparse NID alignment, i.e. only choosing a subset of pixels to calculate NID, has not been tried before, but would be a direct extension of this work. Second, We would like to use image pyramid in the future. We hope to further imporve the computation speed and the scale robustness. Finally, we'd like to develop tracking and mapping algorithm based on the current framework.

TABLE IV

SUCCESSFUL ESTIMATES FOR TUM-RGBD DATASET

	DIR	DSO	NID4
fr1/desk	8.89%	10.89%	64.44%
fr1/teddy	15.12%	6.56%	57.78%
fr1/360	5.89%	22.22%	75.56%
fr2/desk	5.89%	9.55%	46.67%
fr2/360_kidnap	12.78%	9.67%	26.67%
fr2/360_pioneer	9.67%	10.12%	40.00%
fr3/long_house	5.22%	21.56%	86.67%
fr3/str_texture_far	16.67%	13.56%	42.22%
fr3/str_texture_near	5.56%	7.78%	37.78%

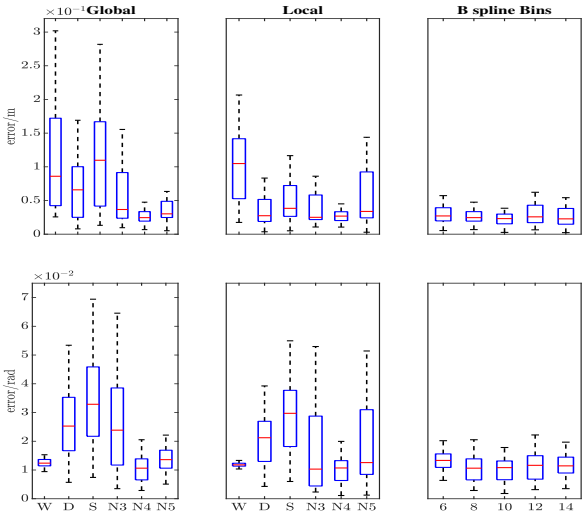

Fig. 5. Median/max/min translation error for both the direct method and different splits of the NID algorithm for the ETHZ-CVG illumination change dataset. The first row is the translation error, the second row is the rotation error. For the column, from left to the right, the figures are results from global/local illumination change datasets and different B-spline bin numbers. **W**: Our simple direct method. **D**: DSO. **S**: SVO. **N*k***: NID method with k splits. Red line: median value. The bottom and top edges of the box: 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles, respectively. The whiskers: extreme data and outliers.

VII. ACKNOWLEDGEMENTS

We thank Juan Falquez for his initial investigations on this topic. This material is based upon work supported by the National Science Foundation under Grant No. 1764092.

REFERENCES

[1] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, (Washington, DC, USA), IEEE Computer Society, 2011.

[2] S. Klose, P. Heise, and A. Knoll, "Efficient Compositional Approaches for Real-Time Robust Direct Visual Odometry from RGB-D Data," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, November 2013.

[3] W. Churchill and P. Newman, "Experience-based navigation for long-term localisation," *The International Journal of Robotics Research*, vol. 32, no. 14, pp. 1645–1661, 2013.

[4] M. J. Milford and G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pp. 1643–1649, IEEE, 2012.

[5] L. Di Stefano, S. Mattoccia, and F. Tombari, "Zncc-based template matching using bounded partial correlation," *Pattern recognition letters*, vol. 26, no. 14, pp. 2129–2134, 2005.

[6] S. Mattoccia, F. Tombari, and L. Di Stefano, "Reliable rejection of mismatching candidates for efficient zncc template matching," in *2008 15th IEEE International Conference on Image Processing*, pp. 849–852, IEEE, 2008.

[7] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[8] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *European conference on computer vision*, pp. 151–158, Springer, 1994.

[9] F. Stein, "Efficient computation of optical flow using the census transform," in *Joint Pattern Recognition Symposium*, pp. 79–86, Springer, 2004.

[10] H. Alismail, B. Browning, and S. Lucey, "Bit-planes: Dense subpixel alignment of binary descriptors," *arXiv preprint arXiv:1602.00307*, 2016.

[11] L. Clement and J. Kelly, "How to train a cat: Learning canonical appearance transformations for direct visual localization under illumination change," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2447–2454, 2018.

[12] X. Wu and C. Pradalier, "Illumination robust monocular direct visual odometry for outdoor environment mapping," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 2392–2398, IEEE, 2019.

[13] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, p. 1, 2017.

[14] M. Kasper, F. Nobre, C. Heckman, and N. Keivan, "Unsupervised metric relocalization using transform consistency loss," in *Conference on Robot Learning (CoRL)*, 2020.

[15] M. Li, X. Chen, X. Li, B. Ma, and P. M. Vitányi, "The similarity metric," *IEEE transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[16] A. Stewart, *Localisation using the Appearance of Prior Structure*. PhD thesis, University of Oxford - New College, 2014.

[17] F. Maes, D. Vandermeulen, and P. Suetens, "Medical image registration using mutual information," *Proceedings of the IEEE*, vol. 91, no. 10, pp. 1699–1722, 2003.

[18] P. Viola and W. M. Wells III, "Alignment by maximization of mutual information," *International journal of computer vision*, vol. 24, no. 2, pp. 137–154, 1997.

[19] B. Delabarre and E. Marchand, "Camera localization using mutual information-based multiplane tracking," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1620–1625, IEEE, 2013.

[20] G. Pascoe, W. P. Maddern, and P. Newman, "Robust direct visual localisation using normalised information distance.," in *BMVC*, pp. 70–1, 2015.

[21] G. Pascoe, W. Maddern, M. Tanner, P. Piniés, and P. Newman, "Nid-slam: Robust monocular slam using normalised information distance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1435–1444, 2017.

[22] G. Caron, A. Dame, and E. Marchand, "Direct model based visual tracking and pose estimation using mutual information," *Image and Vision Computing*, vol. 32, no. 1, pp. 54–63, 2014.

[23] A. Dame and E. Marchand, "Second-order optimization of mutual information for real-time image registration," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4190–4203, 2012.

[24] K. Pathak, A. Birk, N. Vaskevicius, M. Pfingsthorn, S. Schwertfeger, and J. Poppinga, "Online three-dimensional slam by registration of large planar surface segments and closed-form pose-graph relaxation," *Journal of Field Robotics*, vol. 27, no. 1, pp. 52–84, 2010.

[25] L. Carlone and G. C. Calafiore, "Convex relaxations for pose graph optimization with outliers," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 1160–1167, 2018.

[26] S. B. Lazarus, I. Ashokaraj, A. Tsourdos, R. Zbikowski, P. M. Silson, N. Aouf, and B. A. White, "Vehicle localization using sensors data fusion via integration of covariance intersection and interval analysis," *IEEE Sensors Journal*, vol. 9, no. 7, pp. 1302–1314, 2007.

[27] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[28] J. P. Pluim, J. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: a survey," *IEEE transactions on medical imaging*, vol. 22, no. 8, pp. 986–1004, 2003.

[29] S. Park, T. Schöps, and M. Pollefeys, "Illumination change robustness in direct visual slam," in *ICRA*, 2017.

[30] S. Baker and I. Matthews, "Lucas-kanade 20 years on: A unifying framework," *International journal of computer vision*, vol. 56, no. 3, pp. 221–255, 2004.

[31] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in *Numerical analysis*, pp. 105–116, Springer, 1978.

[32] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g 2 o: A general framework for graph optimization," in *2011 IEEE International Conference on Robotics and Automation*, pp. 3607–3613, IEEE, 2011.

[33] J.-L. Blanco, "A tutorial on se (3) transformation parameterizations and on-manifold optimization," *University of Malaga, Tech. Rep*, vol. 3, p. 6, 2010.

[34] E. Eade, "Lie groups for 2d and 3d transformations," *URL http://ethaneade. com/lie. pdf, revised Dec*, vol. 117, p. 118, 2013.

[35] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 3, pp. 611–625, 2017.

[36] C. O. Daub, R. Steuer, J. Selbig, and S. Kloska, "Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data," *BMC bioinformatics*, vol. 5, no. 1, p. 118, 2004.

[37] A. Venelli, "Efficient entropy estimation for mutual information analysis using b-splines," in *IFIP International Workshop on Information Security Theory and Practices*, pp. 17–30, Springer, 2010.

[38] A. Perperoglou, W. Sauerbrei, M. Abrahamowicz, and M. Schmid, "A review of spline function procedures in r," *BMC medical research methodology*, vol. 19, no. 1, p. 46, 2019.

[39] S. Park, T. Schöps, and M. Pollefeys, "Illumination change robustness in direct visual slam," in *2017 IEEE international conference on robotics and automation (ICRA)*, pp. 4523–4530, IEEE, 2017.

[40] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, Oct. 2012.

[41] J. Engel, T. Schöps, and D. Cremers, "Lsd-slam: Large-scale direct monocular slam," in *European conference on computer vision*, pp. 834–849, Springer, 2014.