

Multimodal Colored Point Cloud to Image Alignment

Noam Rotstein Amit Bracha Ron Kimmel
Technion - Israel Institute of Technology
{snoamr, amit.bracha, ron}@cs.technion.ac.il

Abstract

Reconstruction of geometric structures from images using supervised learning suffers from limited available amount of accurate data. One type of such data is accurate real-world RGB-D images. A major challenge in acquiring such ground truth data is the accurate alignment between RGB images and the point cloud measured by a depth scanner. To overcome this difficulty, we consider a differential optimization method that aligns a colored point cloud with a given color image through iterative geometric and color matching. In the proposed framework, the optimization minimizes the photometric difference between the colors of the point cloud and the corresponding colors of the image pixels. Unlike other methods that try to reduce this photometric error, we analyze the computation of the gradient on the image plane and propose a different direct scheme. We assume that the colors produced by the geometric scanner camera and the color camera sensor are different and therefore characterized by different chromatic acquisition properties. Under these multimodal conditions, we find the transformation between the camera image and the point cloud colors. We alternately optimize for aligning the position of the point cloud and matching the different color spaces. The alignments produced by the proposed method are demonstrated on both synthetic data with quantitative evaluation and real scenes with qualitative results.

1. Introduction

In recent years, research in 3D shape reconstruction has made tremendous progress. Much of the research in the field focused on shape reconstruction from IR or RGB images. Multiple approaches have been proposed including amongst others, shape from stereo and monocular depth estimation. More recently, the focus of the community shifted towards supervised deep learning methods [9, 12, 13, 19, 31, 44, 48, 59, 60]. Deep Learning relies heavily on large and accurate datasets. Since such datasets are difficult to obtain, most works use synthetic datasets to train

and validate their models. However, these datasets are limited since they do not capture real-world properties such as distortions and noise. In addition, the use of deep learning methods may require specific training data for each camera model, as the shape reconstruction algorithms are sensitive to model properties and artifacts.

One solution to acquire accurate ground-truth depth is to use precise, yet, often slow, depth measuring devices such as 3D laser scanners. Using such a device to acquire exact depth values for a desired camera model requires registering the device pose relative to the camera one. Placing the device at a fixed and calibrated position relative to the camera is often unsuitable for this task. Such a setup suffers from technical difficulties and requires constant maintenance [20]. Accurate laser scanners usually require a long scanning time for a given scene. Consequently, in order to create a large-scale dataset, multiple images should be acquired for each depth scan. This means that a 3D Euclidean transformation must be found between each image coordinate system and the scanner coordinate system.

The geometric information and color texture provided by some of these devices can be combined to produce a colored point cloud that can be used for registration. However, such registration faces the challenge of multimodality. More specifically, with the comparison of two images taken with different devices - one with the scanner and one with the camera. Different devices have different color properties, often referred to as color gamut. When two devices capture the same scene, the captured color space of each device is different. This defines the problem that our work attempts to solve. How to accurately align a colored point cloud with a colored image under multimodal conditions.

The multimodal alignment task, which is critical to produce accurate and reliable geometric data, has not yet been directly explored. While a rough alignment can be easily found, a precise alignment is required for a proper geometric dataset construction. Therefore, in this paper, we focus on refining such an alignment. Direct visual odometry methods can be modified to be used for our task. Such methods have been mainly used in SLAM, Optical Flow and Color Mapping. These approaches attempt to minimize

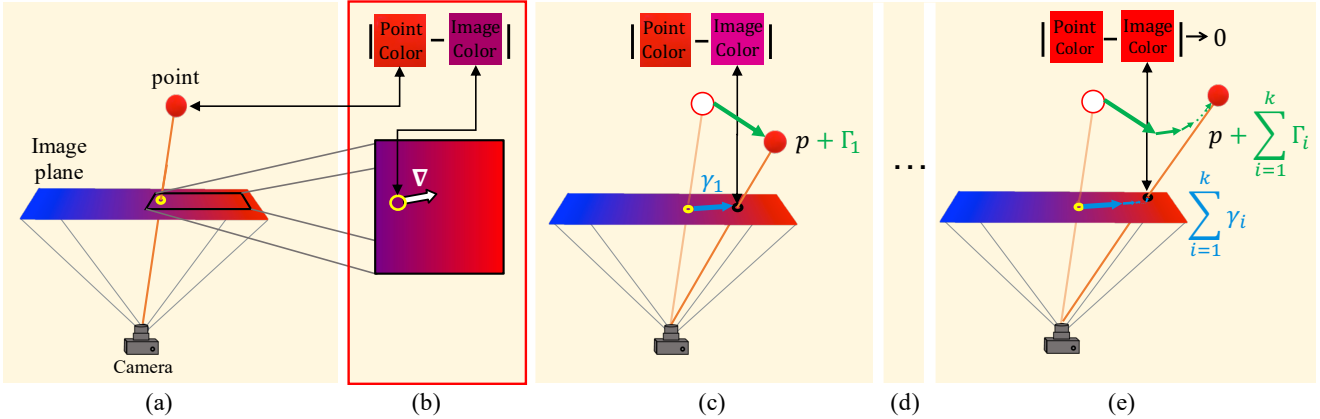


Figure 1. The proposed framework operating, for simplicity, on a single point under a single modality setting. (a) Project the point onto the image plane. (b) Examine the photometric difference between the point color and the image color at the projected location. Unlike projection, color on the image plane is not straightforward differentiable. Create a differential surface on the image plane. (c) Use the differentiation to compute the gradient of the translation parameters and perform a step in the direction of the gradient. (d) Move the point iteratively to optimize the photometric difference. (e) Stop when the photometric error between the point color and the image color at the projected point location is minimal. In the case of many points in our point cloud, the update is also applied to the rotation parameters.

the photometric error between corresponding pixels based on the estimated scene geometry. However, unlike our task, these approaches align images acquired by the same device.

In contrast to previous methods, the proposed pipeline uses a direct numerical sub-pixel scheme to approximate the gradients in the image plane. We prove that the commonly used method for computing such gradients is equivalent to evaluating the gradients on a blurred image. To overcome multimodality, we also propose a second-order polynomial color transformation between the point cloud colors and the image colors. Such an approach has rarely been used for color transformations. The proposed approach produces state-of-the-art results for multimodal colored point cloud to image alignment.

2. Related Efforts

Alignment and registration have been studied intensively in many different directions and setups. One of the most fundamental tasks is 3D-to-3D alignment between two point clouds. The most common solution is the Iterative closest point algorithm (ICP) [4, 10]. Many improvements and variations of this algorithm have been proposed [5, 27, 47, 52]. Other studies proposed to align the colored 3D point cloud using the additional RGB color information [26, 38, 50] or hue values [46].

The task of 2D-to-2D image registration has also been widely explored and used in many applications. The most popular approach is to find corresponding points in both images and then determine the transformation between them [11]. The two leading methods for finding corresponding points are intensity-based methods and feature-based meth-

ods. Intensity-based alignment methods compare intensity patterns in images and image patches [21, 49]. Feature-based methods extract features in each image [1, 3, 42, 54] and then match them. Mapping between the image coordinates is then derived from the corresponding matches.

While 3D-to-3D and 2D-to-2D tasks have been thoroughly explored, we address a different challenge. We are interested in fitting the colors of the 3D geometry to the 2D image. Such a procedure [28], aligns 3D point clouds to overhead images using edge costs and free space costs. Visual-based localization (VBL), is a domain that attempts to approximate camera pose relative to known 3D models [51]. The most common method is to use image feature descriptors. Features are extracted on a query image and compared to features coupled with 3D coordinates. Then, registration is performed using the Perspective-n-Point (PnP) algorithm [40, 43]. In contrast to our task, which focuses on accurate registration, most VBL papers and benchmarks focus on efficient and fast matching between image features and features of large-scale geometric models [18, 41, 55].

There are some approaches that can be associated with the proposed method for pose optimization. Zhou and Koltun proposed to align multiple images to an uncolored point cloud [61]. As opposed to the method we present, they require a few images, optimize a large number of parameters to find the colors of the points, and operate under a single modality setup. Pulli et al. align two colored point clouds by minimizing color and range on two image planes [53], also under single modality assumptions. These methods can be classified as direct visual odometry (DVO). They optimize the geometry directly on the image intensities by minimizing the photometric reprojection error be-

tween images. These techniques are used for camera localization [15, 33, 57], simultaneous localization and mapping (SLAM) with RGB-D cameras [2, 32], SLAM with stereo cameras [17], and SLAM with monocular cameras [16]. Some of the algorithms used for SLAM estimate the geometric model to perform registration. Thus, although these algorithms attempt to register 2D images, their process can be related to our task. A key difference between SLAM and the problem addressed is the multimodal configuration of the former. In addition, we use a different direct numerical gradient approximation scheme that leads to a significant improvement in the alignment. Other SLAM methods attempt to match between features in images in conjunction with their 3D coordinates [6]. Unlike these methods, which optimize the geometric model while aligning images, we benefit from an accurate geometric model that can be used in our favor.

Some DVO methods attempt to perform affine lighting corrections [17] or optimize gamma correction [15] while performing alignment. However, these methods first convert the RGB values to grayscale. We aim to operate in a multimodal environment and compare color values from different devices. This comparison requires a color manipulation. Such manipulations have been widely explored, but have not yet been used for pose estimation. The problem can be viewed as the gamut mapping problem, where the task is to find a transformation of color images from input to output devices. Examples of such solutions are space-dependent gamut mappings [37, 45]. Sochen et al. show how different models of color perception, interpreted as geometries of the color space, lead to different enhanced processing schemes [56]. Much of the work in this area focuses on the perceptual relationship between colors rather than their precise values. For our concern, these solutions are not appropriate, since our problem requires the quantitative comparison of their values. A classical approach to color manipulation of images is histogram equalization. Caselles et al. try to overcome the fact that histogram modification sometimes does not produce good contrast by performing it locally on connected components of the image [8]. These methods succeed in improving image contrast, but were not designed for analytical comparison of color values. Specifically, they do not consider color relations between corresponding pixels in different images. Many papers attempt to perform color matching for value comparison when converting RGB signals to standard CIE tristimulus values. Typical methods include three-dimensional lookup tables with interpolation [25], neural networks [30], and polynomial regression models [29]. Lookup tables lack the differentiability necessary for our task. In the method we propose, the corresponding colors to be matched are computed per iteration, so training a neural network each time is not a feasible solution. On the contrary, poly-

nomial regression models satisfy the necessary requirements for our goal. Several experiments have investigated the influence of polynomial order on the success of color transformation [24, 58]. From these experiments, it can be concluded that the higher the order used, the better the results. Practically, second order models proved to provide accurate transformations at low computational cost.

3. Rigid Alignment and Color Matching

In this section, we show how to directly align a colored point cloud to the perspective image of a given scene. We assume a pinhole camera model with known intrinsic parameters. Our model takes advantage of the fact that nearby pixels in natural images tend to have similar colors and that the color change is slow and gradual. We leverage this property in an optimization scheme by moving the point cloud so that its colors and the colors of the image on the projected points locations match. For simplicity, we first consider a case where the colors of the point cloud and the colors of the image were captured by the same device and share the same color gamut. We denote the point cloud $\{x_j\}_{j=1}^n \in \mathbb{R}^{n \times 3}$ in the XYZ space, and its colors $\{c_j\}_{j=1}^n \in \mathbb{R}^{n \times 3}$, in the RGB space. The 2D coordinates of a point projected onto the image plane are denoted by $p_j \in \mathbb{R}^2$. The colors of the image $I : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ at p_j are given by $I(p_j)$. Using the former notations, when a point is aligned, $c_j \approx I(p_j)$. Note that unlike the image plane, the perspective projection of the point cloud onto the plane is a straightforward differentiable operation. To obtain a fully differentiable procedure, we define a differentiable surface. This surface is discussed and analyzed in Section 3.2. In the proposed procedure, the discrepancy depends on the 3D point coordinates. For a one point scenario, we compute a 3D translation $\Gamma \in \mathbb{R}^3$ where the coordinates of the point in \mathbb{R}^3 are $x_j + \Gamma$. Thus, the discrepancy is differentiable by Γ . We use an iterative optimization procedure to find the translation Γ that minimizes $|I(p_j + \gamma) - c_j|$ where $\gamma \in \mathbb{R}^2$ is the projected location of $x_j + \Gamma$ on the plane (see Fig 1).

Given a point cloud containing n points, we repeat this operation for each point and move the point cloud as a whole to minimize the total color difference of the points. To translate the points, we apply a 3D Euclidean transformation, namely translation and rotation. The last transformation, denoted T^θ , has six parameters ($\theta \in \mathbb{R}^6$) that are iteratively updated during the optimization process.

In the former simple case described, we assumed that the point cloud and image had identical color gamut. This assumption generally does not hold and images and point clouds captured with two distinct cameras sensing the same scene have different color values. To get a meaningful and accurate alignment, we need to compensate for such color discrepancies between the two different types of sensors we are using. Various DVO methods convert colors to

grayscale and work with a single color channel. In contrast, we propose to use the three-dimensional RGB space. We attempt to compensate for color discrepancies without prior color manipulation or color calibration. At each iteration, we have a correspondence between the colors of the point cloud and the colors of the image. That is, we have $3 \times n$ values to match. One effective solution is to use this correspondence to find a linear transformation $D^{\text{linear}} \in \mathbb{R}^{3 \times 4}$ in the 3D color space. This transformation fits between the two sets of colors. However, in practice, a linear mapping of the colors cannot capture the complexity of color discrepancies [24] (see Fig 3). To this end, we propose the linear transformation $D \in \mathbb{R}^{3 \times 10}$ that minimizes the difference between second-order polynomials from one set of RGB colors to the other. Although choosing a higher order polynomial provides higher accuracy, the marginal error between the second and higher order is small [58]. Therefore, using second order provides the required accuracy while maintaining computational efficiency.

3.1. Sub-Pixel Color

A digital RGB image can be viewed as a discrete function

$$J(a, b) \in [0, 1]^3 \quad a, b \in \mathbb{Z}. \quad (1)$$

Therefore, sub-pixel $u, v \in \mathbb{R}$ color values require interpolation. We opt for the classical bilinear interpolation (BL),

$$I(u, v) = BL(J)(u, v). \quad (2)$$

3.2. Sub-Pixel Color Gradient

Besides the gradient calculation in the image plane, our pipeline consists of straightforward differential steps. Given the discrete image J , we need to estimate the gradients between the pixel points. Similar to the different definitions of pixel gradients [34], one can also use different definitions for sub-pixel gradients. In this section we study two different definitions for such a calculation. Strategy A , which is used in our method, and Strategy B , which is the the common calculation method in DVO implementations [2, 16, 17, 32, 33, 61].

1. **Strategy A**- The sub-pixel color values $I(u, v)$ can be viewed as a differential function of u, v . To obtain the gradient, this function is differentiated directly by u ,

$$I_u^A(u, v) = I_u(u, v) = BL(J)_u(u, v), \quad (3)$$

and similarly by v .

2. **Strategy B**- First, the central finite difference approximation image of J is computed,

$$J_a(a, b) = \frac{J(a+1, b) - J(a-1, b)}{2}, \quad (4)$$

and similarly for b . Then, the sub-pixel accuracy is computed, by bilinear interpolation of the discrete gradient images,

$$I_u^B(u, v) = BL(J_a)(u, v), \quad (5)$$

and similarly for v .

Thus, the difference can be explained as follows, In strategy A a continuous representation of the image is constructed and then differentiated, while in strategy B a discrete gradient image is computed and then interpolated. Let us analyze the two strategies using a 1-D example with linear interpolation of a discrete function $h(a)$, $a \in \mathbb{Z}$. The sub-integer continuous interpolated values of the function at $x \in \mathbb{R}$ are,

$$f(x) = (1 - \delta) \cdot h(x_j) + \delta \cdot h(x_{j+1}). \quad (6)$$

Where f is the continuous function estimate, $x_j = \lfloor x \rfloor$, $x_{j+1} = \lceil x \rceil$ and $\delta = x - \lfloor x \rfloor$. In strategy A , as shown in Equation (3), the differentiation is done directly by x and thus by δ ,

$$f_x^A(x) = h(x_{j+1}) - h(x_j) \triangleq \Delta h_j. \quad (7)$$

Let us examine strategy B . First, the gradients of the discrete function are computed according to Equation (4),

$$h_a(a) = \frac{h(a+1) - h(a-1)}{2}. \quad (8)$$

Similar to Equation (5), linear interpolation is used to calculate the sub-pixel gradient:

$$\begin{aligned} f_x^B(x) &= (1 - \delta) \cdot h_a(x_j) + \delta \cdot h_a(x_{j+1}) = \\ &= \frac{(1 - \delta) \cdot \Delta h_{j-1} + \Delta h_j + \delta \cdot \Delta h_{j+1}}{2}. \end{aligned} \quad (9)$$

The proof of the last transition and an extension to the 2D image domain is provided in the supplementary material. Equations (7, 9) can be used to relate the two strategies by the following convolution,

$$f_x^B(x) = f_x^A * w(x). \quad (10)$$

Where w is a rectangular window function,

$$w(x) = \begin{cases} 0.5, & -1 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

We conclude that the popular strategy B for gradient computation is actually a smoothed version of the gradient computed by strategy A . It is equivalent to applying the gradient proposed in strategy A to a blurred image. This, in turn, implies loss of high frequency information and gradients that are affected by values of distant pixels. This simple, yet crucial, distinction between strategy A and the commonly used

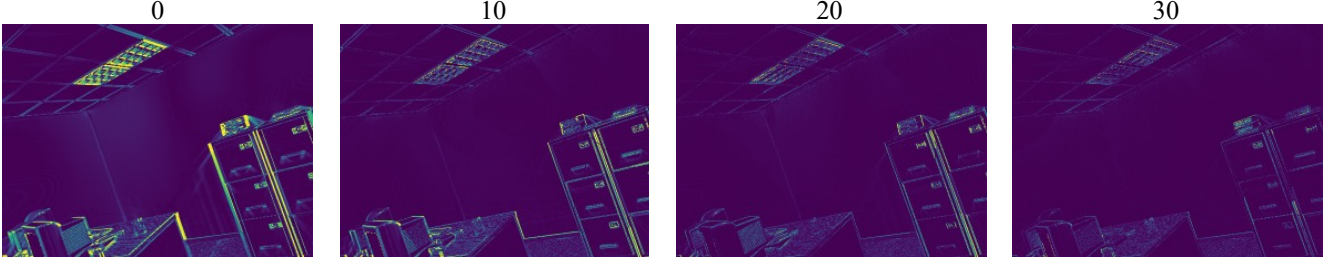


Figure 2. A visualization of the color difference during optimization iterations between an image and an image rendered from the point cloud. Blue corresponds to small values, while yellow corresponds to large values. One can see a significant difference in the initial non-aligned phase and the decrease of the difference as the optimization progresses.

strategy B has a significant impact on the accuracy of the alignment. The ablation study in Section 4.3 demonstrates this important observation both empirically and quantitatively.

Our numerical approximation preferred strategy can be viewed from a different perspective. The comparison between strategy A and strategy B relates to the difference between central finite difference approximations and forward or backward approximations. See, for example, [34]. The reason for choosing central difference approximations lies in a numerical error evaluation derived from a truncated Taylor expansion. The approximation is relevant assuming that $\Delta x \ll 1$, where Δx is the sampling interval of the continuous image function. However, if the information in the image involves high frequencies, this assumption may be misleading. In such a case, the approximation error in deriving the numerical approximations of the derivatives is of an order of the change in the function they approximate. We argue that high frequencies are crucial for accurate alignment and therefore tight numerical stencils are better suited for the task. In most of our scenarios, we could assume that $\Delta x \gg 1$, in which the approximation would fail to properly capture the relevant numerical error. For this particular case, there are better options than the central differentiation strategy B . Indeed, a preferred option would be a tight numerical stencil that uses only x_j and x_{j+1} as in strategy A .

3.3. Color Transformation

The colors of the point cloud c and the corresponding interpolated image colors I^p of its points p are obtained from different camera sensors. Therefore, to compare them, we would like to find the proper color relation between them. We assume that we can write each color of the image as a function of the colors of the corresponding point in the point cloud. To approximate this unknown function, we apply a second-order polynomial kernel to the colors $\{I^p\}$,

$$\begin{aligned} K(I^p) &= K(R^p, G^p, B^p) \\ &= [1, R, G, B, RG, GB, RB, R^2, G^2, B^2] \in \mathbb{R}^{10 \times n}. \end{aligned} \quad (12)$$

In contrast to the framework of Hong et al. [24] for camera colorimetric characterization, we do not add the 3rd order

term RGB as an additional dimension. The reason is that the experimental results have shown no significant advantage when this dimension is added. The point cloud alignment improves in each iteration. Therefore, the correspondence between the color values of the point cloud and the color values of the image improves as well. To exploit this, we find the color transformation repeatedly for each iteration i . Then, the transformation is applied to compute the transformed colors derived from the image

$$I^{D_i} = D_i K(I^p). \quad (13)$$

3.3.1 Color Transformation Optimization

To avoid outliers affecting the color transformation, a scheme for affine illumination correction [17] is used. An inlier point is defined as a point that holds,

$$\|I_j^{D_i} - c_j\| < \beta_{max}. \quad (14)$$

The series of coefficients $D_i \in \mathbb{R}^{3 \times 10}$ of the polynomial terms that minimizes the sum of the color differences of the inliers is computed by the least squares method,

$$D_i = \arg_D \min \|I - c\|^2. \quad (15)$$

The combined optimization problem for finding the inliers and computing D_i is solved alternatively and iteratively. In the supplementary, we show how to handle color values exceeding $[0, 1]$ while preserving differentiability.

3.4. Proposed Scheme

The complete procedure for each iteration i can be described by the following steps.

1. Transform the coordinates x with the current transformation T^{θ_i} .
2. Apply a Z-buffer E to mask out occluded points.
3. Project the 3D coordinates onto the image plane using perspective projection Proj.

4. Calculate the interpolated image colors of the projected points using I .
5. Lift interpolated image colors with a second order polynomial kernel K .
6. Find the color transformation D_i between kernel applied colors and point cloud colors and transform colors.

The computed color values depend on θ_i ,

$$I^{*i} = D_i K (I (\text{Proj} (E (T^{\theta_i}(x)))))) \in \mathbb{R}^{3 \times n}. \quad (16)$$

3.5. Alignment Optimization

We can define the photometric error for each point and for each color channel l ,

$$r_{jl} = I_{jl}^{*i} - c_{jl}. \quad (17)$$

Using the above definitions, the loss function can be calculated. A weighted nonlinear least squares approach is chosen,

$$L(\theta_i) = \|I^{*i} - c\|_W^2 = \sum_{j=1}^n \sum_{l=1}^3 w_{jl} r_{jl}^2. \quad (18)$$

The weights w_{jl} are computed according to the proposed non-Gaussian error models [33], which assume t-distributed errors with $\nu = 5$ degrees of freedom,

$$w_{jl} = \frac{\nu + 1}{\nu + \frac{r_{jl}^2}{\sigma^2}}. \quad (19)$$

The value of σ is computed iteratively,

$$\sigma = \frac{1}{3n} \sum_{j=1}^n \sum_{l=1}^3 r_{jl}^2 \frac{\nu + 1}{\nu + \frac{r_{jl}^2}{\sigma^2}}. \quad (20)$$

We implemented and optimized the proposed algorithm with Pytorch and Pytorch Autograd and used the Adam algorithm for optimization. Since the angle and translation parameters are of different units and orders of magnitude, we initialize their learning rate accordingly. For a megapixel image, the method converges in a few hundred iterations (see Fig 2), which takes about 30 seconds on a single GeForce GTX 2080ti GPU. Since runtime or memory constraints are not critical to dataset construction, we do not focus on an efficient implementation.

4. Experiment setups

In this section we show how the proposed method performs in two scenarios. The first, involves a synthetic database. By controlling the image generation process, we

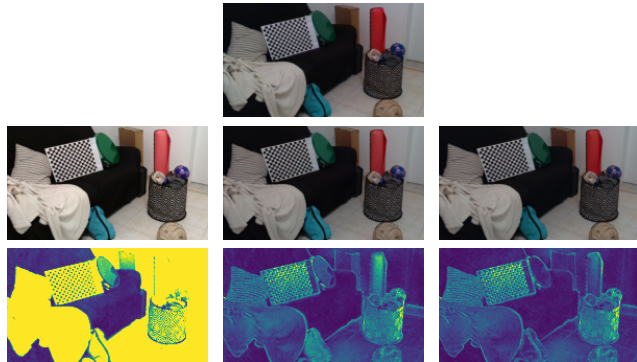


Figure 3. Top row - an RGB image captured by an Intel RealSense Depth Camera D435. Middle row - a rendered image from a point cloud scanned with a FARO 3D Focus Laser Scanner (left), the rendered image after a linear color transformation (middle), and the rendered image after a second-order polynomial transformation (right). The bottom row shows the absolute difference between the transformed modified image and the original image. Blue corresponds to small values, while yellow corresponds to large values. One can see that the color discrepancies of the linear transformation are reduced when using the second order polynomial transformation.

can accurately quantify the accuracy of the method. Next, we apply the proposed method to real images. We demonstrate success in aligning a FARO laser scanner point cloud with an image captured by an RGB camera. The alignment is evaluated qualitatively by projecting the intensity edges of the rendered point cloud onto the edges of the given image.

4.1. Synthetic Data

To demonstrate the proposed method, a photorealistic dataset is needed. We use the synthetic dataset ICL-NUIM [22], which contains significant synthetic depth and RGB noise models representing a realistic environment. From the dataset, 2000 images and their corresponding consecutive images are randomly sampled. A corresponding point cloud of the images is created using their corresponding depth values. These point clouds are misaligned with the successive images. To simulate multimodality, we apply a series of effects to each of the RGB images. Many effects can be applied, such as effects that take into account spatial color configuration [37] or perform retinex modifications [14, 36, 39]. To simplify the experiment, a set of elementary effects was chosen. Their numerical values with an example are discussed in the supplementary material,

1. Apply random color transformation of brightness, contrast, saturation, and hue.
2. Apply gamma correction with a random gamma value.

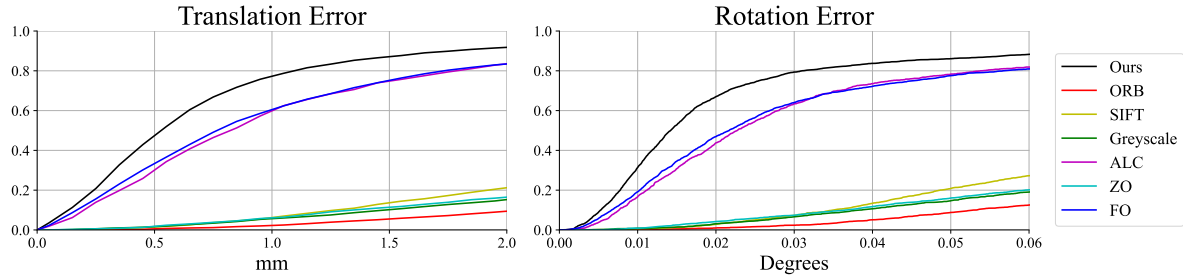


Figure 4. Cumulative normalized histograms of translation and rotation errors for synthetic data experiment. The proposed method with second-order polynomial color alignment outperforms the rest of the methods.

3. Simulate different point spread functions and sensor properties by applying a Gaussian blur to the image.

Point cloud to image alignment has no direct comparison. Therefore, we study the use of modifications of different works,

1. **ORB** - Feature-based camera localization with Oriented FAST and Rotated BRIEF (ORB) [54] features is one of the most popular approaches for SLAM [6]. This approach is based on frame-to-frame registration in conjunction with coupled depth values. To modify it for our task, an image is rendered from the point cloud. ORB features are found on the original image and the rendered image. The point cloud 3D coordinates are used to determine the 3D coordinates of the features found on the rendered image. After matching the feature descriptors between the images, the Ransac algorithm is applied together with the PnP algorithm to compute the Euclidean transformation.
2. **SIFT** - A common methodology in VBL implementations is to use Root Sift features [1, 42]. After the features are found and matched, the alignment is performed as in using ORB features.
3. **Greyscale** -DVO methods commonly use grayscale images [33, 57] instead of RGB images. In this setup, we convert the colors to grayscale and then perform the alignment.
4. **Affine Lighting Correction (ALC)** - Engel et al. [17] propose to first convert the colors to grayscale. Then, an affine lighting Correction is performed by alternately optimizing two parameters. These parameters form an affine intensity transform that corrects the grayscale values.
5. **Zero Order (ZO)** - Our pipeline used with no color transformation with direct comparison of the RGB color values.

6. **First Order (FO)** - Our pipeline used with first-order color transformation instead of second order.

It is important to note that methods 3, 4, 5, 6 have been modified to use our scheme for computing gradients. This scheme is tested separately in Section 4.3.

The error is captured by two measures, the translation error and the rotation error. The translation error is the Euclidean distance between the original translation and the derived translation. To calculate the rotation error, the combined rotation axis is found using the rotation error about each axis. The error is calculated by computing the rotation about each axis. We present the cumulative normalized histogram of these errors under the different configurations in Fig. 4. Our approach clearly outperforms the other configurations. Linear color transformations (FO and ALC), while beneficial, are inferior to the second-order color transformation. Feature-based approaches (ORB and SIFT) and approaches that do not perform color transformation (ZO and Greyscale) achieve worse results compared to the alternatives. We speculate that accurate geometry leads DVO methods with color modifications to significantly surpass feature-based approaches. We denote that with the ICL-NUIM dataset and our method, we achieve a median sub-millimeter translation error of 0.58mm and a median rotation error of 0.014°.

4.2. Real Data

To show that the proposed method works on real data, we used a FARO 3D Focus Laser Scanner and an Intel RealSense Depth Camera D435. We only use the RGB image from this camera and not its depth sensing capabilities. After each scan is completed, we place a camera near the scanner position. As mentioned earlier, the method relies on a coarse estimate of the camera pose. We place a checkerboard in the scene and use it to roughly estimate the camera pose relative to the scanner. This will be the initial transformation we use before applying the proposed method.

For simplicity, we use a checkerboard pattern. Such an initialization can also be found using one of the methods described in Section 2. For example, by applying a

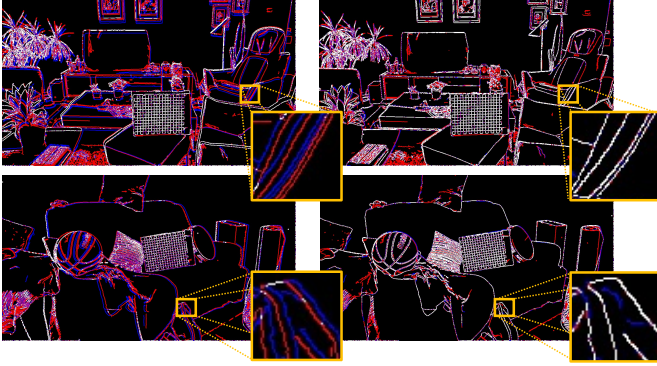


Figure 5. A comparison between edges from camera images and edges extracted from rendered point cloud images. Left is at initialization, right - proposed refinement. The colors denote the origin of the edges: red- camera image, blue- rendered image from point cloud, white- overlap. In the initial rough alignment, the corresponding edges do not overlap.

coarse-to-fine scheme [57]. Unlike the scenario with the synthetic data, we do not have the ground truth transformation. Therefore, we estimate the success of the proposed method by visually comparing the RGB image and the rendered image from the point cloud after applying the computed transformation. Since different cameras are used, a simple color difference is not a good visual measure. Image edges are less affected by the camera characteristics and reflect whether the images are aligned correctly. Therefore, we find the edges of each image using the Canny-Haralick edge detector [7, 23, 35] and compare the edge images. Fig. 5, shows the edges extracted from both rendered and camera images, see the supplementary material for more examples. Due to the imperfect extraction of edges, not all edges in each image are detected. We can see that the edges in both images are aligned. As we can observe, this is in contrast to the edge comparison of the first misalignment of the edges. Although we cannot quantify the exact error in the real data scenario, the edge representation shows how accurate the proposed method is.

4.3. Ablation Study

We demonstrate the impact of our sub-pixel gradient computation using the synthetic experiment setup. We perform a test of our method with a single difference. The calculation of the sub-gradient is changed. The alternative configuration uses the usual gradient calculation in DVO implementations as described in Section 3.2. From the experimental results (Fig. 6), such a configuration leads to inferior alignment compared to our method. The median translation error increases by 13.2% and the median rotation error increases by 18.4%.

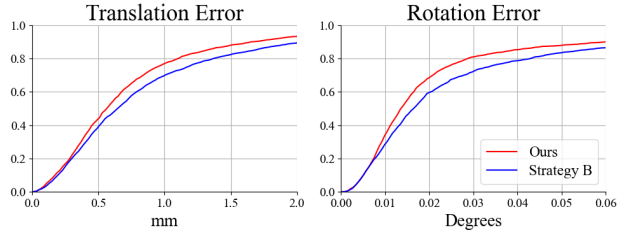


Figure 6. Cumulative normalized histograms of translation and rotation errors for the ablation study experiment. The proposed method for computing sub-pixel gradients on the image plane outperforms the common approach.

5. Limitations

As shown, our sub-pixel gradient computation benefits from high-frequency detail and improves alignment accuracy in the ICL-NUIM dataset. Although the dataset contains a simulation of real world noise, we speculate that extremely noisy scenarios may benefit from blurring and loss of such detail. However, we believe that such blurring should be intentional. The presented method may suffer from additional limitations common to DVO methods. For example, if the camera is distant from the scanner, the point cloud may contain missing parts that do not appear on the image, or vice versa. This could potentially degrade the alignment results.

6. Conclusions

We introduced an iterative differential method that aligns a colored point cloud to an image in a multimodal environment using geometric and second-order polynomial color matching and gradient-based optimization. The proposed framework introduces an algorithmic pipeline that uses the entire point cloud and image information to minimize the discrepancy between the point cloud colors and projected image colors. We analyze the computation of the gradient on the image plane and show an efficient and direct form of computing it. We explain and numerically support the advantages of using second-order polynomials for color transformation between different camera devices. We believe that the proposed concepts could facilitate and improve the creation of real 3D datasets in the future and could be applied to any camera model.

Acknowledgements- This work addresses a challenge we were presented with during an internship at Intel RealSense. We would like to thank Hila Eliyahu Grosman and Aviad Zabatani for their assistance. The research was supported by the D. Dan and Betty Kahn Michigan-Israel Partnership for Research and Education, run by the Technion Autonomous Systems and Robotics Program.

References

- [1] Relja Arandjelović and Andrew Zisserman. Three things everyone should know to improve object retrieval. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2911–2918. IEEE, 2012. 2, 7
- [2] Cedric Audras, A Comport, Maxime Meilland, and Patrick Rives. Real-time dense appearance-based slam for rgb-d sensors. In *Australasian Conf. on Robotics and Automation*, volume 2, pages 2–2, 2011. 3, 4
- [3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 2
- [4] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992. 2
- [5] Alexander Bronstein, Michael Bronstein, and Ron Kimmel. In the rigid kingdom. In *Numerical Geometry of Non-Rigid Shapes*, pages 119–135. Springer, 2008. 2
- [6] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual-inertial, and multi-map slam. *IEEE Transactions on Robotics*, 2021. 3, 7
- [7] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 8
- [8] V. Caselles, J.-L. Lisani, J.-M. Morel, and G. Sapiro. Shape preserving local histogram modification. *IEEE Transactions on Image Processing*, 8(2):220–230, 1999. 3
- [9] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018. 1
- [10] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992. 2
- [11] Suma Dawn, Vikas Saxena, and Bhudev Sharma. Remote sensing image registration techniques: A survey. In *International Conference on Image and Signal Processing*, pages 103–112. Springer, 2010. 2
- [12] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 1
- [13] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *arXiv preprint arXiv:1406.2283*, 2014. 1
- [14] Michael Elad, Ron Kimmel, Doron Shaked, and Renato Keshet. Reduced complexity retinex algorithm via the variational approach. *Journal of visual communication and image representation*, 14(4):369–388, 2003. 6
- [15] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 3
- [16] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 3, 4
- [17] Jakob Engel, Jörg Stückler, and Daniel Cremers. Large-scale direct slam with stereo cameras. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1935–1942. IEEE, 2015. 3, 4, 5, 7
- [18] Mengdan Feng, Sixing Hu, Marcelo H Ang, and Gim Hee Lee. 2d3d-matchnet: Learning to match keypoints across 2d image and 3d point cloud. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4790–4796. IEEE, 2019. 2
- [19] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2011, 2018. 1
- [20] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1
- [21] Guy Godin, Marc Rioux, and Rejean Baribeau. Three-dimensional registration using range and intensity information. In *Videometrics III*, volume 2350, pages 279–290. International Society for Optics and Photonics, 1994. 2
- [22] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *2014 IEEE international conference on Robotics and automation (ICRA)*, pages 1524–1531. IEEE, 2014. 6
- [23] Robert M Haralick. Digital step edges from zero crossing of second directional derivatives. *IEEE transactions on pattern analysis and machine intelligence*, 6(1):58–68, 1984. 8
- [24] Guowei Hong, M Ronnier Luo, and Peter A Rhodes. A study of digital camera colorimetric characterization based on polynomial modeling. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 26(1):76–84, 2001. 3, 4, 5
- [25] Po-Chieh Hung. Colorimetric calibration in electronic imaging devices using a look-up-table model and interpolations. *Journal of Electronic imaging*, 2(1):53–61, 1993. 3
- [26] Andrew Edie Johnson and Sing Bing Kang. Registration and integration of textured 3d data. *Image and vision computing*, 17(2):135–147, 1999. 2
- [27] Ibrahim Jubran, Alaa Maalouf, Ron Kimmel, and Dan Feldman. Provably approximated icp. *arXiv preprint arXiv:2101.03588*, 2021. 2
- [28] Ryan S Kaminsky, Noah Snaveley, Steven M Seitz, and Richard Szeliski. Alignment of 3d point clouds to overhead images using edges and free space cost. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 63–70. IEEE, 2009. 2

- [29] Henry R Kang. Color scanner calibration. *Journal of imaging science and technology*, 36(2):162–170, 1992. 3
- [30] Henry R Kang and Peter G Anderson. Neural network applications to the color scanner and printer calibrations. *Journal of Electronic Imaging*, 1(2):125–135, 1992. 3
- [31] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017. 1
- [32] Christian Kerl, Jürgen Sturm, and Daniel Cremers. Dense visual slam for rgb-d cameras. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2100–2106. IEEE, 2013. 3, 4
- [33] Christian Kerl, Jürgen Sturm, and Daniel Cremers. Robust odometry estimation for rgb-d cameras. In *2013 IEEE international conference on robotics and automation*, pages 3748–3754. IEEE, 2013. 3, 4, 6, 7
- [34] Ron Kimmel. The level set method: Numerical considerations. *Numerical Geometry of Images*, pages 61–74, 2004. 4, 5
- [35] Ron Kimmel and Alfred M Bruckstein. Regularized laplacian zero crossings as optimal edge integrators. *International Journal of Computer Vision*, 53(3):225–243, 2003. 8
- [36] Ron Kimmel, Michael Elad, Doron Shaked, Renato Keshet, and Irwin Sobel. A variational framework for retinex. *International Journal of computer vision*, 52(1):7–23, 2003. 6
- [37] Ron Kimmel, Doron Shaked, Michael Elad, and Irwin Sobel. Space-dependent color gamut mapping: A variational approach. *IEEE Transactions on image processing*, 14(6):796–803, 2005. 3, 6
- [38] Michael Korn, Martin Holzkothen, and Josef Pauli. Color supported generalized-icp. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 3, pages 592–599. IEEE, 2014. 2
- [39] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971. 6
- [40] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. *International journal of computer vision*, 81(2):155, 2009. 2
- [41] Liu Liu, Hongdong Li, and Yuchao Dai. Efficient global 2d-3d matching for camera localization in a large-scale 3d map. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2372–2381, 2017. 2
- [42] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2, 7
- [43] C-P Lu, Gregory D Hager, and Eric Mjolsness. Fast and globally convergent pose estimation from video images. *IEEE transactions on pattern analysis and machine intelligence*, 22(6):610–622, 2000. 2
- [44] Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5695–5703, 2016. 1
- [45] John McCann. Lessons learned from mondrians applied to real images and color gamuts. In *Color and imaging conference*, volume 1999, pages 1–8. Society for Imaging Science and Technology, 1999. 3
- [46] Hao Men, Biruk Gebre, and Kishore Pochiraju. Color point cloud registration with 4d icp algorithm. In *2011 IEEE International Conference on Robotics and Automation*, pages 1511–1516. IEEE, 2011. 2
- [47] Niloy J Mitra, Natasha Gelfand, Helmut Pottmann, and Leonidas Guibas. Registration of point cloud data from a geometric optimization perspective. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, pages 22–31, 2004. 2
- [48] Jiahao Pang, Wenxiu Sun, Jimmy SJ Ren, Chengxi Yang, and Qiong Yan. Cascade residual learning: A two-stage convolutional neural network for stereo matching. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 887–895, 2017. 1
- [49] Xenophon Papademetris, Andrea P Jackowski, Robert T Schultz, Lawrence H Staib, and James S Duncan. Integrated intensity and point-feature nonrigid registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 763–770. Springer, 2004. 2
- [50] Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Colored point cloud registration revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 143–152, 2017. 2
- [51] Nathan Piasco, Désiré Sidibé, Cédric Demonceaux, and Valérie Gouet-Brunet. A survey on visual-based localization: On the benefit of heterogeneous data. *Pattern Recognition*, 74:90–109, 2018. 2
- [52] François Pomerleau, Francis Colas, Roland Siegwart, and Stéphane Magnenat. Comparing icp variants on real-world data sets. *Autonomous Robots*, 34(3):133–148, 2013. 2
- [53] Kari Pulli, Simo Piiroinen, Tom Duchamp, and Werner Stuetzle. Projective surface matching of colored 3d scans. In *Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05)*, pages 531–538. IEEE, 2005. 2
- [54] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2, 7
- [55] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *2011 International Conference on Computer Vision*, pages 667–674. IEEE, 2011. 2
- [56] Nir Sochen, Ron Kimmel, and Ravi Malladi. A general framework for low level vision. *IEEE transactions on image processing*, 7(3):310–318, 1998. 3
- [57] Frank Steinbrücker, Jürgen Sturm, and Daniel Cremers. Real-time visual odometry from dense rgb-d images. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, pages 719–722. IEEE, 2011. 3, 7, 8
- [58] Ibrahim Yilmaz, I Oztug Bildirici, Murat Yakar, and Ferruh Yildiz. Color calibration of scanners using polynomial transformation. In *XXth ISPRS Congress Commission V, Istanbul, Turkey*, pages 890–896, 2004. 3, 4

- [59] Jure Zbontar, Yann LeCun, et al. Stereo matching by training a convolutional neural network to compare image patches. *J. Mach. Learn. Res.*, 17(1):2287–2318, 2016. [1](#)
- [60] Feihu Zhang, Victor Prisacariu, Ruigang Yang, and Philip HS Torr. Ga-net: Guided aggregation net for end-to-end stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 185–194, 2019. [1](#)
- [61] Qian-Yi Zhou and Vladlen Koltun. Color map optimization for 3d reconstruction with consumer depth cameras. *ACM Transactions on Graphics (TOG)*, 33(4):1–10, 2014. [2](#), [4](#)