Translational Photometric Alignment of Single-View Image Sequences

Adam P. Harrison\textsuperscript{a,*}, Dileepan Joseph\textsuperscript{a}

\textsuperscript{a}Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, T6G-2V4

Abstract

Photometric stereo is a well-established method to estimate surface normals of an object. When coupled with depth-map estimation, it can be used to reconstruct an object’s height field. Typically, photometric stereo requires an image sequence of an object under the same viewpoint but with differing illumination directions. One crucial assumption of this configuration is perfect pixel correspondence across images in the sequence. While this assumption is often satisfied, certain setups are susceptible to translational errors or misalignments across images. Current methods to align image sequences were not designed specifically for single-view photometric stereo. Thus, they either struggle to account for changing illumination across images, require training sets, or are overly complex for these conditions. However, the unique nature of single-view photometric stereo allows one to model misaligned image sequences using the underlying image formation model and a set of trans-
lational shifts. This paper introduces such a technique, entitled translational photometric alignment, that employs the Lambertian model of image formation. This reduces the alignment problem to minimizing a nonlinear sum-squared error function in order to best reconcile the observed images with the generative model. Thus, the end goal of translational photometric alignment is not only to align image sequences, but also to produce the best surface-normal estimates given the observed images. Controlled experiments on the Yale Face Database B demonstrate the high accuracy of translational photometric alignment. The utility and benefits of the technique are further illustrated by additional experiments on image sequences suffering from uncontrolled real-world misalignments.

Keywords: Image alignment, photometric stereo, nonlinear minimization, maximum likelihood estimation, depth map reconstruction

1. Introduction

Photometric stereo, introduced by Woodham [1], is a well-established method to calculate surface normals of an object. When followed by depth-map estimation, photometric stereo is also a crucial step in reconstructing object depth maps. Typically, photometric stereo is performed using a sequence of images of an object with known and controlled illumination all under the same viewpoint. When employing single-view photometric stereo, it is assumed that every image in the sequence shares perfect pixel correspondence.

Normally, this poses no problems, as photometric stereo is typically applied using a fixed camera setup in environments stable enough to ensure
adequate pixel correspondence. However, certain image capture setups may not guarantee pixel correspondence between images. In particular, image frames may be corrupted by relative translational misalignments.

Important examples of such scenarios include performing photometric stereo with reflected light [2] or scanning electron microscopy [3, 4]. Apart from microscopy, it may be difficult in other imaging setups to completely fix the object of interest. For instance, as this paper will demonstrate, when capturing faces a human subject may move his or her head enough to affect photometric stereo performance.

The changing pixel intensities across light directions in an image sequence make it challenging to correct for translational misalignments. However, the distinct nature of single-view photometric stereo image sequences offers a specific means of alignment. Nonetheless, as image alignment and changing pose is a well-researched area, the scope of this problem overlaps with important areas of related work.

In the past decade, published works have introduced powerful Multi-View Photometric Stereo techniques [5, 6, 7, 8]. As these works are designed for image sequences taken with a camera whose position varies in world coordinates over time, they have none of the pixel correspondence requirements of traditional single-view photometric stereo. Nevertheless, while these techniques are very powerful, they are also unavoidably complex as they are designed for highly challenging image-capture scenarios where the viewpoint is intentionally changed. This complexity is not needed for image capture scenarios that attempt to use the same viewpoint, but suffer from translational errors in alignment. Instead, translational misalignments can
be modelled in image space, avoiding much of the complexity inherent in these techniques.

**Silhouette-Based Alignment** is a common approach to image registration. These techniques are often based on global criteria, such as moments. When performing translational alignment, centering images using silhouette centroids is the appropriate moment-based method to use [9]. This is called *centroid alignment*. However, global silhouette-based methods are sensitive to errors or changes in the silhouette boundary [10]. This poses a problem with photometric-stereo image sequences, as boundaries of extracted silhouettes may differ at every illumination direction due to differences in pixel intensities. Other more robust alternatives use methods that attempt to model noise and occlusions into the silhouette description. These methods can be based on representations describing the boundary or the entire silhouette [11]. Unfortunately, when using non global silhouette-based methods, it is not clear how to best evaluate the results [11]; thus, there is no agreed upon way to determine when error-prone silhouettes have been successfully aligned. Finally, both global and non-global silhouette matching require an object mask, which is not always possible to generate.

**Intensity-Based Alignment** methods involve selecting features between a target and source image, determining correspondences between these features, and finally computing a transform aligning the target image with the source [12]. Typical features include corners, lines, and image patches. Two very well known examples of registration techniques using image patches as features are cross-correlation [13] and optical flow [14]. Regardless of the features used for alignment, intensity-based techniques are all based in some way
upon pixel values in the images. As a result, traditional image patch-based methods will struggle when aligning photometric stereo image sequences, and the changing intensity levels across images make consistent and reliable detection of lower-level features, such as lines or corners, a much more difficult problem. These reasons motivate the use of features outside of traditional intensity-based ones.

Along those lines, Appearance-Varying Alignment enhancements to the optical flow procedure have provided a means to match intensity values under varying illumination [15, 16]. Yet, aside from simple variations such as gain and offset, these techniques require computing basis images using a training set of the object. This is not necessarily a drawback for tracking applications, which these techniques are designed in part to address. In fact, computing basis images beforehand can often be desirable for real-time tracking, as it delegates a significant amount of computing to an offline task. However, training sets are not practical for aligning photometric-stereo image sequences, especially for large datasets of objects.

In the context of photometric stereo, a better option is to use Model-Based Alignment, which employs basis images derived from models of image formation. Armed with such a model, no training set is required. When imaging objects with diffuse reflectance characteristics, the Lambertian model of image formation is well-established for relating pixel intensities to surface normals and illumination direction. Thus, by incorporating the Lambertian model of image formation, one can exploit intensity-based information in the alignment technique. This paper describes such an image alignment technique. Called translational photometric alignment (TPA), this
novel alignment technique is designed to align single-view images of an object illuminated from differing directions.

2. Method

The foundation of TPA is a model explaining the generation of a misaligned image sequence. Section 2.1 develops such a model, whose framework incorporates Lambertian reflectance and translational misalignments. Using this model, the problem of estimating misalignments corrupting image sequences is reduced to minimizing a nonlinear function. The practical steps required to minimize this nonlinear function are detailed in Section 2.2.

2.1. Modelling Misalignments

As the illumination direction changes from image to image in a sequence, to develop an alignment routine one can incorporate a model of image formation into the process. Image formation models can vary in complexity depending on the properties of the object surface and light sources. When estimating surface normals of diffuse objects, often it is assumed that the object follows Lambertian reflectance and is illuminated by a directional light source. In addition to these conditions, if orthographic cameras, a linear relationship between pixel intensity and irradiance, and no inter-reflections or cast shadows are also assumed, the pixel intensity at image coordinates \((x, y)\) can be expressed as:

\[
I(x, y) = \alpha(x, y)\ell^T\eta(x, y) + \epsilon(x, y),
\]

\[
\alpha(x, y) = u(\ell^T\eta(x, y)),
\]

\[
\eta(x, y) = \rho(x, y)n(x, y),
\]
where \( \mathbf{n}(x, y) = (n_x, n_y, n_z) \) represents surface normals, \( \rho(x, y) \) is the surface albedo, \( \ell \) is the light direction expressed as a unit vector, and \( u(.) \) is the unit step function. The vector \( \mathbf{\eta}(x, y) \) represents the surface normals multiplied by the albedo, and will be referred to as \textit{weighted normals}. Attached shadows are dealt with by the unit step function in (2), which are a crucial and important phenomena [17]. Noise in (1), represented by \( \epsilon(x, y) \), is modelled as independent and identically distributed (IID) additive zero-mean Gaussian noise, a model consistent with camera behaviour [13].

Any violations of the Lambertian assumptions contribute to errors in the model. If these errors become large enough, different image formation and noise models may be more appropriate and should be used instead of (1). For instance, robust photometric stereo methods use the Lambertian model, but weight image observations in some manner, such as with binary values [18] or with ‘soft’ weights [19]. Alternatively, other techniques use specular image formation models, such as the Torrance-Sparrow [20, 21] or Ward [22] model. This is by no means an exhaustive list and, in principle, TPA can accommodate any reflectance model. However, TPA is developed in this paper using the Lambertian model. As experiments will demonstrate, the Lambertian model is effective enough to align image sequences of objects with challenging reflectance properties.

Given an object with a continuous set of weighted normals, a sequence of \( N \) images, and a one-to-one correspondence between horizontal and vertical coordinates of the real-world and pixel-space, TPA uses the following equation to model the formation of a \textit{misaligned} image \( (I_k) \):

\[
I_k(x, y) = \alpha_k(x + \Delta x_k, y + \Delta y_k) \ell_k^T \mathbf{\eta}(x + \Delta x_k, y + \Delta y_k) + \epsilon_k(x, y),
\] (4)
where \(1 \leq k \leq N\). In other words, as defined by (4), misaligned images are individual observations of an object with a continuous set of normals under a set of relative shifts. An image sequence’s set of shifts can be expressed as a single shift vector: \(\mathbf{\beta} = (\Delta x_1, \Delta y_1, \ldots, \Delta x_N, \Delta y_N)\). By way of (4), determining shift values requires simultaneously determining the continuous surface normals. Since image noise follows IID Gaussian distributions, the maximum-likelihood (ML) estimate of \(\mathbf{\beta}\) and \(\eta\) is one that minimizes the sum-squared error (SSE) of all observations [23]. This can be expressed as:

\[
SSE(\mathbf{\beta}, \eta) = \sum_{k=1}^{N} \sum_{x=1}^{n} \sum_{y=1}^{m} r_k(\mathbf{\beta}, \eta, x, y)^2,
\]

where each individual residual term is defined as:

\[
r_k(\mathbf{\beta}, \eta, x, y) = I_k(x, y) - \alpha_k(x + \Delta x_k, y + \Delta y_k)\ell_k^T\eta(x + \Delta x_k, y + \Delta y_k),
\]

and images are assumed to consist of \(m \times n\) pixels.

While (5) may provide a theoretically valid condition on maximizing likelihood, the equation does not offer an avenue in which to easily determine image shifts. Fortunately, a simple change of variables provides a more practical formulation. Substituting \(u = x + \Delta x_k\) and \(v = y + \Delta y_k\), with \(\Delta u_k = -\Delta x_k\) and \(\Delta v_k = -\Delta y_k\), misalignment can be modelled as a set image shifts rather than a set of object shifts. This redefines \(\mathbf{\beta}\) as \((\Delta u_1, \Delta v_1, \ldots, \Delta u_N, \Delta v_N)\), meaning that the image formation equation is reexpressed as:

\[
I_k(u + \Delta u_k, v + \Delta v_k) = \alpha_k(u, v)\ell_k^T\eta(u, v) + \epsilon_k(u + \Delta u_k, v + \Delta v_k).
\]

Implied within this change of variables is an important distinction from (4); namely, images and their accompanying noise are treated as continuous
entities and surface normal values are only considered at a discrete set of locations. Although an image is of course discrete by its nature, it can also be viewed as a continuous function sampled at pixel locations. Interpolating values at locations between pixels approximates the continuous underlying image, thereby providing a means to treat a discrete image as a continuous entity.

On the other hand, being for the most part a discrete phenomenon arising from processes within pixels, there is no continuous analogue for image noise. As a result, there is no appropriate noise distribution at locations in between pixels. Thus, interpolating at location \((u + \Delta u_k, v + \Delta v_k)\) results in a weighted sum of normally-distributed error terms. While a weighted sum of normally-distributed variables is also normally distributed, the distributions of adjacent sums may not be IID. This means that the ML estimate requires a generalized least-squares (GLS) formulation [23]. However, for the sake of simplicity, image noise shall be assumed to be IID and normally-distributed for all locations, even those residing in between pixels.

By modelling misalignments as sets of image shifts, (7) does not account for pixel values of border regions “uncovered” by image shifts. One way to account for these regions is to assume background pixels of each image are of a constant value. Pixel values of border regions uncovered by shifts may then be filled in using the constant background value. Should each image be segmented in order to isolate foreground regions prior to TPA, then the image sequence will implicitly satisfy this condition. For image sequences with no clear background, one can simply restrict the area of interest to a window within each image. Uncovered pixels can then be determined using
pixels outside the window area.

With misalignment modelled as a set of image shifts, the current shift estimate completely determines the current weighted-normal estimate, $\hat{\eta}$. When modelling attached shadows, the maximum likelihood estimate of the weighted normals at each pixel location is the solution to a least squares problem using a reduced regressor matrix [24]:

$$\hat{\eta}(\beta, u, v) = L(\beta, u, v)^+ i(\beta, u, v),$$

where $(.)^+$ denotes the Moore-Penrose pseudo inverse and the regressor matrix is similar to the regressor matrix used in traditional photometric stereo, except that light directions corresponding to attached shadows have been zeroed:

$$L(\beta, u, v) = \begin{pmatrix} \alpha_1(u,v) \ell_1^T \\ \vdots \\ \alpha_N(u,v) \ell_N^T \end{pmatrix}.$$ (10)

In order to estimate weighted normals, $L(\beta, u, v)$ must have a rank of 3. Determining which light directions are zeroed is a difficult problem, as it must be done simultaneously with estimating weighted normals. This paper employs a fixed-point iterative method [24] to estimate weighted normals and determine $L(\beta, u, v)$ for each pixel location.

As a result, rather than simultaneously solving for image shifts and surface normals, only the correct image shift vector $\beta$ need be estimated. This simplifies the sum-squared error formulation, as the SSE is now only a func-
tion of the shift vector $\beta$:

$$SSE(\beta) = r(\beta)^T r(\beta). \quad (11)$$

Here the residuals have been flattened into a single vector as defined by:

$$r(\beta) = (r(\beta, 1, 1)^T \ldots r(\beta, n, m)^T)^T, \quad (12)$$

with the residual vector at each pixel location defined as:

$$r(\beta, u, v) = i(\beta, u, v) - L(\beta, u, v) \hat{\eta}(\beta, u, v), \quad (13)$$

$$= (I - P(\beta, u, v)) i(\beta, u, v), \quad (14)$$

where $P(\beta, u, v) = L(\beta, u, v)L^+(\beta, u, v)$ is the idempotent and symmetric projection matrix onto the range of $L(\beta, u, v)$. Note that entries in (13) differ from (6) by incorporating the current weighted-normal estimate, $\hat{\eta}$. Despite this difference, minimizing (11) is equivalent to minimizing (5) if one disregards the issues arising from modelling image noise as continuous phenomena.

2.2. Estimating Misalignments

By minimizing the SSE one determines the ML estimate of the shift vectors. This forms the basis of TPA. This is in effect a nonlinear function minimization scheme, where the only independent variables are the image shifts $\beta$. A great benefit of constructing residuals in this manner is that minimization also produces the best weighted-normal estimates, $\hat{\eta}$, under the photometric stereo assumptions. As a result, the ultimate goal of performing image alignment in the first place is incorporated directly into the residual term.
When evaluating the nonlinear function at continuous locations, image pixel values at these locations must be approximated using an interpolation method. In this paper, bilinear interpolation is used to approximate image values at these continuous locations.

Minimizing the nonlinear SSE function of (11) can be accomplished using appropriate function minimization schemes. Many minimization techniques require gradient evaluations, which in turn require partial derivative computations of the SSE term with respect to each of the variables in the shift vector $\mathbf{\beta}$. These variables can represent shifts in either the $u$ or $v$ directions for one particular image. For instance, one can consider the partial derivative of (11) with respect to shifts in the $u$ direction of $I_k$. This partial derivative can be evaluated separately at each pixel location:

$$\frac{\partial \text{SSE}(\mathbf{\beta}, u, v)}{2\partial \Delta u_k} = \frac{\partial \mathbf{r}(\mathbf{\beta}, u, v)}{\partial \Delta u_k} \mathbf{r}(\mathbf{\beta}, u, v).$$

(15)

To compute (15), the partial derivative of (14) with respect to $\Delta u_k$ must be evaluated. This can be expressed as:

$$\frac{\partial \mathbf{r}(\mathbf{\beta}, u, v)}{\partial \Delta u_k} = \mathbf{Q}(\mathbf{\beta}, u, v) \frac{\partial \mathbf{i}(\mathbf{\beta}, u, v)}{\partial \Delta u_k} - \frac{\partial \mathbf{P}(\mathbf{\beta}, u, v)}{\partial \Delta u_k} \mathbf{i}(\mathbf{\beta}, u, v),$$

(16)

where $\mathbf{Q}(\mathbf{\beta}, u, v) = \mathbf{I} - \mathbf{P}(\mathbf{\beta}, u, v)$ is the idempotent and symmetric complement of $\mathbf{P}(\mathbf{\beta}, u, v)$. Due to the discontinuous nature of the unit step function incorporated into $\mathbf{P}(\mathbf{\beta}, u, v)$ and the coupled nature between weighted-normal estimates and the $\mathbf{L}(\mathbf{\beta}, u, v)$ matrix, the derivative of the projector matrix is undefined. However, if one assumes that the set of light directions used to estimate $\mathbf{\eta}(\mathbf{\beta}, u, v)$ will remain constant for the interval in question, then the expression in (16) simplifies to:

$$\frac{\partial \mathbf{r}(\mathbf{\beta}, u, v)}{\partial \Delta u_k} = \mathbf{Q}(\mathbf{\beta}, u, v) \frac{\partial \mathbf{i}(\mathbf{\beta}, u, v)}{\partial \Delta u_k},$$

(17)
The partial derivative of the $\mathbf{i}(\mathbf{\beta}, u, v)$ vector has all zero entries except for the $k^{th}$ entry, which will evaluate as the spatial derivative of $I_k(\mathbf{\beta}, u, v)$ in the $u$ direction. With the simplification from (16) to (17), the problem is equivalent to a general separable nonlinear least squares (GSNLS) problem [23]. This is desirable, as the GSNLS structure of the problem can be used to simplify gradient evaluations [25]. Substituting (17) into (15) to complete the gradient expression results in:

$$\frac{\partial \text{SSE}(\mathbf{\beta}, u, v)}{2 \partial \Delta u_k} = \frac{\partial \mathbf{i}(\mathbf{\beta}, u, v)}{\partial \Delta u_k}^T \mathbf{Q}(\mathbf{\beta}, u, v) \mathbf{r}(\mathbf{\beta}, u, v),$$

(18)

$$= \frac{\partial \mathbf{i}(\mathbf{\beta}, u, v)}{\partial \Delta u_k} \mathbf{r}(\mathbf{\beta}, u, v),$$

(19)

$$= \frac{\partial I_k(\mathbf{\beta}, u, v)}{\partial \Delta u_k} r_k(\mathbf{\beta}, u, v),$$

(20)

where (19) uses the identity $\mathbf{P}(\mathbf{\beta}, u, v) \mathbf{r}(\mathbf{\beta}, u, v) = 0$ and (20) takes advantage of the sparse nature of the partial derivative of $\mathbf{i}(\mathbf{\beta}, u, v)$. The complete partial derivative of the SSE can be evaluated by computing and adding together (20) for every pixel location and multiplying the result by 2. An identical approach is used for evaluating the gradients with respect to shifts in the $v$ direction. As the spatial image gradients play such a crucial role, it is important to evaluate them with high accuracy. This paper calculates the spatial gradient using a centered-difference formula of $O(h^6)$ accuracy.

The nature of GSNLS problems also provides a specific means to compute or approximate the Hessian [25]. Often, explicitly evaluating the Hessian for nonlinear problems is avoided due to the prohibitive cost of evaluating the matrix [23]. However, for the TPA problem the sparsity inherent in the partial derivatives of $\mathbf{i}(\mathbf{\beta}, u, v)$ allow an efficient computation of the complete Hessian. As with the gradient, the Hessian can be evaluated separately at
each pixel location. For instance, the following second-order partial derivative may be derived:

\[
\frac{\partial^2 \text{SSE}(\beta, u, v)}{\partial \Delta u_k \partial \Delta v_\ell} = \frac{\partial I_k(\beta, u, v)}{\partial \Delta u_k} Q_{k\ell}(\beta, u, v) \frac{\partial I_\ell(\beta, u, v)}{\partial \Delta v_\ell} + \delta_{k\ell} \frac{\partial^2 I_k(\beta, u, v)}{\partial \Delta u_k \partial \Delta v_k} r_k(\beta, u, v),
\]

(21)

where the subscripts on \( Q(\beta, u, v) \) denote specific entries and the \( \delta_{k\ell} \) represents the Kronecker delta function. Thus, the second term in (21) is only nonzero for second-order partial derivatives of shifts of the same image. Similar expressions exist for second-order partial derivatives in solely the \( u \) or \( v \) directions. To compute the complete Hessian matrix, the expression in (21) must be evaluated and summed across all pixel locations and for all combinations of shift parameters.

With expressions for the gradient and Hessian developed, established minimization routines can be employed to determine the shift vector. Trust-region methods are a popular and effective variation of classic Newton minimization and have performed very well in many applications [23]. However, since the SSE function is nonconvex, minimization routines can be susceptible to local minima, the potential of which increases with larger misalignments relative to image dimensions. If this is an issue with the images at hand, more recent and specialized minimization techniques may be required. For instance, coarse-to-fine minimization techniques, whose theory has been explored for the optical flow problem [26], may also prove useful for the TPA problem. Nonetheless, as experiments in this paper will demonstrate, trust-region minimization can successfully correct significant misalignments.

The above formulation only describes relative misalignments between im-
ages in a sequence. As a result, every image in a sequence can be shifted by the same amount without affecting the SSE. To avoid this, TPA anchors the first image of the sequence and aligns the remaining images.

3. Results

This section outlines the results of two experiments that explore the performance and benefits of TPA. The first experiment, discussed in Section 3.1, tests the ability of TPA to correct for known translations on the Yale Face Database B [27]. This allows the performance of TPA to be judged under a controlled but challenging scenario. On the other hand, Section 3.2 tests whether TPA can provide tangible benefits to a real-world scenario suffering from uncontrolled translational misalignments. All experiments employed trust-region minimization using the interior reflective technique [28, 29] implemented in MATLAB.

Stopping criteria consisted of relative tolerances of $10^{-4}$ for both the objective function and the shift values, which are two orders of magnitude higher than MATLAB's default values. These higher values were used to speed up convergence because, once step sizes and SSE values become small enough, inaccuracies inherent to image interpolation will begin to dominate all other factors. Experience indicates that, provided small enough values are given, TPA is not overly sensitive to these tolerances. Thus, identical tolerances were used for all image sequences in the experiments.

The two sets of experiments measure the benefits that TPA provides to photometric stereo, a key motivation of TPA. The coefficient of determination, or $R^2$ value, is used to quantify improvements. In its most general form,
$R^2$ values measure the improvement of one hypothesis over another, and can be expressed as:

$$R^2 = \frac{SSE(H_0) - SSE(H_1)}{SSE(H_0)},$$

where $H_0$ and $H_1$ represent the null and test hypotheses, respectively. In the context of these experiments, weighted normals calculated from misaligned and TPA-corrected images constitute $H_0$ and $H_1$ respectively. The SSE is evaluated using (11).

By calculating $R^2$ values this way, (22) summarizes how much of the residual error between actual and modelled images is derived from misalignments within the image sequences that were corrected by TPA. Measurement errors, noise, uncorrected misalignments, and the inherent limitations of the chosen generative model are responsible for any remaining residual values. An $R^2$ value of 1 would indicate that TPA completely reconciled any difference between the generative model and the observed images and a value of 0 indicates that TPA provided no improvements.

3.1. Controlled Experiments

Controlled experiments were performed on the Yale Face Database B [27], which consists of facial pictures of 10 different subjects under 9 different poses using 64 single-source known illumination conditions. As the 64 images for each pose were taken quickly, subject movement is minimal, providing an excellent dataset on which to apply controlled experiments. Moreover, since the reflectance characteristics of the human face are complex and non-Lambertian [30], the Yale Face Database B provides a highly challenging dataset. Section 3.1.1 explains in more detail the makeup of the dataset used.
Table 1: Light directions used in the controlled experiments. Directions are provided in hemispherical coordinates using the \((\theta, \phi)\) convention, where \(\theta \in [0, 90^\circ]\) and \(\phi \in (-180^\circ, 180^\circ]\) (\(\theta\) represents zenith angle and \(\phi\) represents azimuth).

<table>
<thead>
<tr>
<th>(\theta)</th>
<th>0</th>
<th>11.2</th>
<th>11.2</th>
<th>11.2</th>
<th>11.2</th>
<th>22.3</th>
<th>22.3</th>
<th>22.3</th>
<th>22.3</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\phi)</td>
<td>0</td>
<td>116.7</td>
<td>63.3</td>
<td>-116.7</td>
<td>-63.3</td>
<td>154.5</td>
<td>25.5</td>
<td>-154.5</td>
<td>-25.5</td>
<td>180</td>
</tr>
<tr>
<td>(\theta)</td>
<td>25</td>
<td>37.7</td>
<td>37.7</td>
<td>39.7</td>
<td>39.7</td>
<td>45</td>
<td>50</td>
<td>50</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>(\phi)</td>
<td>0</td>
<td>159.7</td>
<td>20.3</td>
<td>-154.0</td>
<td>-26.0</td>
<td>90</td>
<td>180</td>
<td>0</td>
<td>180</td>
<td>0</td>
</tr>
</tbody>
</table>

for experimentation and also outlines a preprocessing step executed to mitigate misalignments in the original dataset. This is followed by Section 3.1.2, which summarizes experiments and performance results of TPA using known misalignments of varying magnitude.

3.1.1. Yale Face Dataset

Overall, the Yale Face Database B provides 9 poses per subject; however, these experiments only use the frontal-pose image sequences (pose 0 in the dataset). All images were resized to \(240 \times 320\) pixels, which is half the original image dimensions. As well, experiments used a subset of the illumination conditions, consisting of 20 light directions. These are provided in Table 1.

As the dataset provides coordinates of the eyes and mouth of each subject for each image, one can create a window of each image consisting of the subject’s face. Thus, TPA can be executed on the windowed area, with areas outside of the windows used to fill in any pixels uncovered by corrective shifts. For each subject, these experiments used a rectangular window comprising the facial feature coordinates given for the first image of the sequence, but padded by 40 pixels in the \(x\) and \(y\) directions. Window sizes did vary across
different subjects, but all were on the order of $130 \times 130$ pixels. Fig. 1 depicts the facial window of one of the database subjects.

Images in the Yale Face Database B were captured over a two second interval, meaning that change in pose was minimal [27]. Nonetheless, these minimal head movements were significant enough to cause minor misalignments of their own within the original image sequences. For certain of these image sequences, such as Subject 4 and 6, these misalignments are significant enough to be visually apparent when viewing the sequence as a video. As a result, an important preprocessing step prior to experimentation consisted of executing TPA on the original unshifted images. Doing so ensured that experiments used a valid baseline. For each of the 10 image sequences, the alignment routine anchored the first image while applying corrective shifts to the remaining 19 images.

As Table 2 demonstrates, depending on the image sequence, errors in image reconstructions after this preprocessing step were reduced by up to 19%. Table 2 also illustrates that corrective shifts were relatively small in
Table 2: Statistics for $R^2$ values after applying TPA on the original Yale Face Database B image sequences. Subjects are ordered by ascending $R^2$ values and shifts are indicated in pixels.

<table>
<thead>
<tr>
<th>Subject</th>
<th>$R^2$</th>
<th>Min Shift</th>
<th>Med. Shift</th>
<th>Max Shift</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4.23%</td>
<td>0.00</td>
<td>0.41</td>
<td>2.06</td>
</tr>
<tr>
<td>10</td>
<td>4.52%</td>
<td>0.07</td>
<td>0.69</td>
<td>1.89</td>
</tr>
<tr>
<td>8</td>
<td>5.12%</td>
<td>0.05</td>
<td>0.54</td>
<td>1.63</td>
</tr>
<tr>
<td>5</td>
<td>5.59%</td>
<td>0.00</td>
<td>0.56</td>
<td>1.34</td>
</tr>
<tr>
<td>9</td>
<td>5.80%</td>
<td>0.00</td>
<td>0.36</td>
<td>1.87</td>
</tr>
<tr>
<td>3</td>
<td>6.67%</td>
<td>0.02</td>
<td>0.24</td>
<td>1.40</td>
</tr>
<tr>
<td>1</td>
<td>8.38%</td>
<td>0.23</td>
<td>0.78</td>
<td>1.52</td>
</tr>
<tr>
<td>7</td>
<td>8.54%</td>
<td>0.02</td>
<td>0.42</td>
<td>1.90</td>
</tr>
<tr>
<td>4</td>
<td>11.75%</td>
<td>0.11</td>
<td>0.72</td>
<td>3.07</td>
</tr>
<tr>
<td>6</td>
<td>19.06%</td>
<td>0.00</td>
<td>0.15</td>
<td>2.19</td>
</tr>
</tbody>
</table>

magnitude. Even so, image sequences that were particularly jittery in their original form were much less so after alignment. Two such examples are provided by Figs. 2 and 3, which depict original and TPA-corrected video sequences of Subject 4 and 6, respectively.

As well, despite their relatively small magnitude, corrective shifts can have a serious impact in the $R^2$ results. Visually, for those image sequences with high $R^2$ values, reconstructed images from weighted-normal estimates were sharper.
Figure 2: Video clip of Subject 4’s image sequence. The left frame depicts the original sequence and the right frame depicts the same sequence after application of TPA.

Figure 3: Video clip of Subject 6’s image sequence. The left frame depicts the original sequence and the right frame depicts the same sequence after application of TPA.
3.1.2. **Alignment Accuracy and Benefits**

Using the preprocessed image sequences, experiments applied known misalignments to the Yale Face dataset. These misalignments were then corrected by TPA. Since misalignments are known, alignment accuracy can be readily measured. If the routine functioned perfectly, corrective shifts would be identical in magnitude and opposite in direction to misalignment shifts. Residual misalignment (RM), defined as the Euclidean distance of any misalignments left over after execution of the alignment routine, provides a means to measure any discrepancies between actual and corrective shifts. The RM for a single image can be written as:

\[ RM = \sqrt{(\Delta u_{\text{actual}} + \Delta u_{\text{corrective}})^2 + (\Delta v_{\text{actual}} + \Delta v_{\text{corrective}})^2}, \]

where \( \Delta(\cdot)_{\text{actual}} \) and \( \Delta(\cdot)_{\text{corrective}} \) are the actual and corrective shifts respectively. Each consists of components in the \( u \) and \( v \) directions.

All images in the sequences, except for the first image, were misaligned using *uniformly distributed* shifts in the \( u \) and \( v \) directions. Misalignment severity was controlled by varying the range of misalignment distribution. 10 different misalignment distributions were used, comprising maximum magnitudes of 1 to 10 pixels, increasing in increments of 1 pixel. Each of the 10 image sequences were misaligned 5 different times at each of the 10 different distributions. Thus, in total the alignment routine was executed 500 times. As with the preprocessing step, during alignment the first image of the sequence was anchored, with the routine applying corrective shifts to the remaining 19 images.

As each image sequence consists of 19 misaligned images, there were 950 RM values for each misalignment distribution. Fig. 4(a) graphs the RM
for each distribution. As the figure demonstrates, alignment accuracy was very high, with median RM values remaining under 0.1 for all misalignment distributions except when maximum misalignment was 10 pixels. At this maximum misalignment, the median RM value was 0.12. For maximum misalignments in the 1 to 6 pixel range, median and third quartile values stayed relatively constant, remaining under 0.1 and 0.2 pixels, respectively. Compared to the roughly 130 × 130 window sizes used, these RM values are very small. As a result, these results demonstrate that TPA can align the Yale face images with almost no RM in the 1 to 6 pixel maximum misalignment range.

Outside of this range, median and quartile RM values do begin to increase. Even so, this increase is very gradual compared to the magnitude
of misalignments, and accuracy remains high at all levels of severity. This accuracy is especially notable when considering that a misalignment of 10 pixels is very severe relative to the window size. From the perspective of nonlinear least-squares minimization, this means that when initialization is far from the correct solution, TPA can successfully locate the global minima for the Yale Face images, even when employing an established routine such as trust-region minimization. Moreover, TPA can accomplish this even under the challenging and non-Lambertian conditions that the Yale Face Database B images present.

While the results in Fig. 4(a) demonstrate that TPA has excellent accuracy even at high maximum misalignments, they do not quantify improvements to photometric stereo. For this reason, $R^2$ values were also calculated. There were 50 different $R^2$ values for each misalignment distribution, five for each image sequence. As Fig. 4(b) illustrates, $R^2$ values, or the proportion of errors in reconstruction performance due to translational misalignment, were significant at all levels of severity. In fact, even when the maximum misalignment magnitude is only 1 pixel, TPA reduces the reconstruction error by over 15%. At higher levels of misalignment, $R^2$ values begin to level off, reaching a plateau of almost 80%.

The rapid rise of $R^2$ values from 1 to 5 pixels of maximum misalignment demonstrate the importance of correcting for misalignments even when their magnitude is relatively small. In other words, even when misalignments are slight, imperfect pixel correspondence can significantly reduce the quality of surface normals and albedos estimated in the photometric stereo step. Visually, these differences in $R^2$ values directly translate to qualitative
Figure 5: Example image reconstructions before and after alignment. All images correspond to reconstructions of Subject 3 illuminated from a zenith and azimuth angle of 25° and 0° respectively. The top and bottom rows depict reconstructions using weighted normals estimated from misaligned and TPA-corrected image sequences respectively. Left, middle, and right images correspond to maximum misalignments of 1, 3, and 5 pixels respectively.

Apart from estimating surface normals, a common end goal of photomet-
Figure 6: Example depth maps before and after alignment. All depth maps correspond to Subject 3. The top and bottom rows depict depth maps estimated from misaligned and TPA-corrected image sequences respectively. Left, middle, and right images correspond to maximum misalignments of 1, 3, and 5 pixels respectively. Depth maps were estimated using a maximum likelihood method [24]

Differences in visual quality between depth maps estimated from misaligned and TPA-corrected image sequences can also be observed. Fig. 6 depicts depth maps to illustrate these qualitative differences. As the figure demonstrates, just as with image reconstructions, misalignments blur depth-map estimates. This blurring effect increases in severity in proportion to misalignment magnitude. In contrast, depth maps estimated from TPA-corrected image sequences have almost none of these visual degradations.
3.2. Real-World Scenario

While Section 3.1 demonstrated that TPA can perform very well in a challenging but controlled environment, it is important to test the routine on a real-world scenario. A variety of causes may explain why an image capture scenario for single-view photometric stereo does not guarantee pixel correspondence. One illustrative and important example is digital microscopy.

When capturing images of microscopic objects, often the light or electron source of typical microscopes is restricted to a fixed position. In these cases, it is often more cost-effective and practical to incorporate a motorized stage into the microscope setup to physically rotate the object rather than developing a specialized setup with varying light source directions. As it is the objects that are actually rotated rather than the light source, the captured images are at successively increasing angles with respect to the horizontal. Consequently, each image is simply rotated back digitally, so that all images are horizontally aligned. This is demonstrated in Fig. 7.

Typically, motorized x-y-phi stages consist of a rotational element mounted on top of translational elements. As a result, when rotating an object, the axis of rotation will not be centred on the object, changing its real-world \((x, y)\) location. To keep an object exactly in the centre of the field of view requires both rotating the stage by the desired amount and also translating the stage to account for the object’s new real-world \((x, y)\) coordinates.

While the expected \((x, y)\) coordinates can be calculated, they are sensitive to errors in stage calibration, stage repeatability, and also the distance of the object to the axis of rotation. Thus, expected and actual \((x, y)\) coordinates do not match. In contrast, rotational accuracy only relies on one single factor—
Figure 7: Capturing images of microscopic objects for use in photometric stereo. In the first row are images of a marine microfossil rotated physically at successively increasing angles. In the second row are the same images, except they have been rotated digitally so that they are aligned, giving the impression that the fossil is stationary while the light source rotates to successively decreasing angles.

accuracy of the rotational element of the stage, which is often quite high. Thus, images suffer predominantly from translational errors relative to each other. Since pixel sizes are often on the order of micrometres, calibrating to that accuracy is often unrealistic for typical setups.

These errors can be mitigated somewhat by a silhouette alignment technique, such as centroid alignment. However, as the object is illuminated from different directions across the image sequence, the extracted silhouettes will be slightly different for each image. Thus, due to the variability between silhouettes at the boundary, centred images of the same object will not always align properly.
For these reasons, microscopy image sequences, captured as described above, serve as an excellent dataset in which to test the performance of TPA using a real-world scenario. Section 3.2.1 describes the dataset of image sequences used for these experiments. In addition, the section gives statistics on the corrective shifts applied to the dataset by TPA. Section 3.2.2 summarizes the improvements TPA provided to photometric stereo.

3.2.1. Microfossil Dataset

In studying microscopic particles, often a large dataset of specimens is required. For instance, the large-scale study of marine microfossils, such as foraminiferal tests (forams), is often a crucial task for oil exploration [31] and palaeoclimatology [32]. Extracting 3D models of forams using photometric stereo and depth-map estimation could provide key benefits to these fields. For instance, photometric stereo can be employed to create powerful digital representations of forams and other opaque specimens [2].

Image sequences of $320 \times 320$ pixels were captured for 500 forams using an x-y-phi motorized stage combined with centroid alignment. Silhouettes were extracted using a threshold value of 0.04. Images were captured with a light source at a zenith angle of $30^\circ$ and with azimuths ranging from $0^\circ$ to $340^\circ$ in increments of $20^\circ$. As a result, each image sequence consists of 18 images. Fig. 8 illustrates the degree of difference between silhouettes of a typical centroid-aligned foram. As the figure demonstrates, differences between silhouettes can be significant. After image capture and centroid alignment, TPA was applied to the dataset. The first image in each sequence was anchored, with corrective shifts applied to the remaining images.

Since the dataset consists of 500 image sequences, each with 17 aligned
Figure 8: Silhouette differences in a centroid-aligned foram. An image of a foram is displayed in (a). In (b), 18 silhouettes have been captured of images of the foram illuminated from the same zenith angle but with differing azimuths. Silhouettes were created using thresholding. Each silhouette has been added together, and the result has been rescaled so that areas of complete overlap are represented by pure white. Areas encompassing fewer image silhouettes range from black to grey.

images, the TPA routine applied 17,000 individual corrective shifts (2 for each image). Fig. 9 graphs the sample frequency of the corrective shifts. The median corrective shift magnitude was roughly 3.5 pixels. Looking at the sample distribution, it is apparent that 3.5 pixels is a representative value of the corrective shift magnitudes. Compared to the 320 × 320 image dimensions, the median value is relatively small.

Even so, these relatively small shifts are visually apparent when watching sequences as a video. This is illustrated by Fig. 10, which depicts a before-and-after video sequence that demonstrates typical improvements TPA provided to the microfossil dataset. As well, corrective shifts of higher magnitudes still constituted a significant portion of the results. The effects of these higher magnitude shifts are even more visually apparent, especially when comparing images whose corrective shifts were applied in opposing di-
In total, 17,000 corrective shifts were applied to the foram dataset. The median corrective shift was 3.5 pixels. As shift magnitude increases, its probability steadily decreases.

3.2.2. Improvements to Photometric Stereo

Since the dataset is affected by uncontrolled translational misalignments, RM cannot be measured. However, improvements to the photometric stereo step can be readily measured. Fig. 13 graphs the sample frequency of $R^2$ values across the microfossil dataset.

As the figure indicates, TPA reduced reconstruction errors predominantly in the 40% to 60% range. In fact, the median $R^2$ value was 49%. The histogram demonstrates that this median value is representative of the different $R^2$ values across the microfossil dataset.

For the most part, misalignments not corrected by centroid alignment
Figure 10: Typical improvements provided by TPA to the foram dataset. The left and right frames depict original and TPA-corrected video sequences, respectively, of a foram. Visually apparent improvements in alignment are apparent.
Figure 11: Visually apparent shifts corrected by TPA. On the left and right, respectively, are two centroid-aligned and TPA-corrected images of a foram. The top and bottom images are illuminated from two different azimuth angles. Important features have been highlighted in red and blue. Due to imperfect segmentation, the silhouettes suffer from error, and features on the centroid-aligned images do not line-up properly. In contrast, despite the significant silhouette errors, TPA performs much better in aligning these features.

were responsible for much of the difference between pixel values of the actual and reconstructed images. Moreover, even the smallest $R^2$ value remained above 15%. Thus, aligning the foram image sequences reduced photometric stereo error for all image sequences by significant amounts. As a result, in order to produce quality weighted-normal estimates, TPA serves as a crucial step for the foram dataset. This demonstrates that TPA can provide important and tangible benefits to an important real-world application of
Figure 12: An example of dramatic improvements provided by TPA to the foram dataset. The left and right frames depict original and TPA-corrected video sequences, respectively, of a foram. Even though misalignments are very severe in the original sequence, the TPA-corrected sequence is a significant improvement compared to the former.
Figure 13: Histogram of $R^2$ values for weighted normals. This figure graphs the distribution of $R^2$ values and illustrates the reduction in reconstruction error of the weighted-normal estimates provided by TPA. In total, there were 500 $R^2$ values, one for each foram. The majority of weighted-normal $R^2$ values cluster around 40% to 60% reduction in error, with the frequency falling off relatively symmetrically on either side.

photometric stereo.

As Section 3.2.1 demonstrated, this significant improvement in the results of photometric stereo only required corrective shifts that were on median 3.5 pixels in magnitude. This indicates that even small amounts of misalignment corrected by TPA can significantly improve the performance of photometric stereo. This is also supported by the results of the Yale Face dataset experiments in Section 3.1.

4. Discussion

TPA is applicable to any image capture setup susceptible to translational misalignments. In many situations, to automatically obtain images of an
object illuminated from differing directions, it is more feasible to rotate the 
object rather than the light source. This is particularly true when imaging 
microscopic objects. For instance, rotating the particle is often the only op-
tion when executing photometric stereo using scanning electron microscopy 
(SEM), due to the inherent complexities and costs of that modality’s equip-
ment [3]. It can also be more effective to rotate objects in the optical domain, 
especially if a motorized stage is already incorporated into the system to lo-
calize and capture images of large numbers of objects at once. Rotating the 
particle can introduce translational misalignments in both the optical [2] and 

Recent works that apply photometric stereo in the SEM domain have 
approached this problem by using detectors able to provide images whose 
contrast is not dependent on the direction of the electron beam [3, 4]. These 
direction-independent images are used for aligning the direction-dependent 
images used for photometric stereo purposes. However, this approach is only 
possible for SEM microscopes equipped with these detectors. In addition, 
the approach is not available to microscopes in the optical domain. As a 
result, an alternative approach is needed in order to correct for translational 
misalignments. The microscopy domain also poses challenges for hardware 
solutions to the alignment problem, such as placing fiducial markers on the 
specimen slides. Due to the microscopic size of specimens, fiducial markers 
would require lithographic etching or some other analogue. For instance, to 
achieve the correction reported in Section 3.2, fiducial markers would have 
to be accurate to 1 µm. For this reason, a software solution to the alignment 
problem can often be preferable to a hardware approach.
TPA would also prove beneficial to applications outside microscopy. For instance, imaging conditions under industrial settings are often susceptible to vibrations and other unpredictable errors. TPA would be able to correct for any translational misalignments caused by such errors. In addition, as the preprocessing step of the Yale Face Database B experiments demonstrated, misalignments can crop up even under well-controlled imaging conditions. While the latent real-world movements of subject faces may be of any movement type, Section 3.1.1 demonstrated that the simplified translational model is still powerful enough to remove much of the jitter in the face image sequences and improve photometric-stereo performance. It should also be noted that placing fiducial markers are not always desirable when imaging human faces or other sensitive subjects.

The experiments presented in this paper successfully applied TPA to image sequences from two highly different setups, highlighting the broad applicability that TPA enjoys. This is due to the use of a generative model to explain and estimate misalignments, which allows TPA to be readily applied to image sequences of objects from a variety of setups, with no need of a training set or specialized hardware. This is an important strength of TPA.

With future work, misalignments other than translational ones may be incorporated. These may include rotational, affine, and even perspective shifts. As with translational misalignments, multi-view photometric stereo techniques [5, 6, 7, 8] inherently handle all such cases. Yet, these approaches work in a world-coordinate frame in order to accommodate an intentionally changing viewpoint, making them excessive for single-view photometric stereo. In contrast, if images are corrupted by minor variations in the real-
world positions of the object and/or camera across all images in the sequence, it is more computationally efficient to approach the problem using an image alignment framework.

While the TPA technique introduced in this paper employed the Lambertian model of image formation, in principle the underlying framework can incorporate any appropriate model. For objects exhibiting more complex reflectance characteristics, models incorporating specularities or other phenomena can relate pixel intensities to surface orientation. As well, TPA can readily accommodate robust photometric stereo techniques designed to remove outliers. However, in these cases estimating surface normals is a more complex problem. Nonetheless, as the Yale Face and microfossil dataset demonstrated, TPA using Lambertian reflectance can effectively align real-world and challenging image sequences.

5. Conclusion

This paper presented a novel alignment routine called TPA that is designed specifically to correct for translational misalignments in a sequence of photometric stereo images. The routine requires that the object in each image is illuminated by known directions. By incorporating a model of image formation directly into its error term, TPA can align images where aligned pixels have varying intensities. This paper detailed a solution using Lambertian reflectance, but other models are possible. Consequently, a direct result of the routine is that estimated surface normals best correspond with the given images under the assumptions of the reflectance model.

The paper also described experiments testing benefits of TPA in two dif-
ferent scenarios. The first scenario used the Yale Face Database B to conduct controlled experiments on 10 image sequences with known misalignments. Uniformly distributed misalignments were applied to the image sequences at 10 different levels of severity. This was repeated 5 times for each image sequence. Experiments demonstrated that even at a maximum misalignment of 10 pixels, TPA can correct for translational errors with an accuracy that was on median better than 0.13 pixels. Moreover, as misalignments of 10 pixels are large relative to the image dimensions used, these results demonstrate that TPA can accurately correct severe misalignments. Correcting for these misalignments also resulted in improvements in the reconstruction performance of photometric stereo. Even when maximum misalignment was only 1 pixel, TPA reduced reconstruction error by on median 15%. Differences in the visual quality of image reconstructions and depth maps between misaligned and TPA-corrected image sequences support these quantitative results.

In addition, the benefits of applying TPA to a dataset of image sequences obtained in a real-world scenario were quantified. The dataset consisted of 500 microfossil image sequences, each consisting of 18 images of $320 \times 320$ pixels. These image sequences inherently suffer from translational misalignments. With a median $R^2$ value of 49%, TPA successively mitigated significant amounts of the reconstruction error of the weighted-normal estimates, meaning that correctable misalignments accounted on median for close to 50% of the reconstruction error in the non-TPA approach. As a result, estimated normals of the specimen surfaces were significantly more consistent with the given images. Since median corrective shift magnitude was only
3.5 pixels, these results, and those of the Yale Face dataset, demonstrate that even slight misalignments can be significantly detrimental to photometric stereo performance. Thus, these results demonstrate the great worth in applying TPA to image sequences affected by translational misalignments.

6. Acknowledgements

The authors gratefully acknowledge the support of the Natural Sciences and Engineering Research Council (NSERC) of Canada and Alberta Innovates – Technology Futures.


39


[27] A. Georgiades, P. Belhumeur, D. Kriegman, From few to many: Illumination cone models for face recognition under variable lighting and


