Practical Approaches for the Estimation of High Resolution Depth Map and Intensity Field using Photometric cue

Swati Sharma and Manjunath V. Joshi

Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat (382007)
sharma_swati@daiict.ac.in, mv_joshi@daiict.ac.in

Abstract

In this paper, we consider the problem of estimating the high resolution surface gradients, albedo and the intensity field using photometric stereo. In real scenario, while using photometric stereo, observations pertaining to light source positions as well as the image intensity values can be noisy. Assuming a Lambertian model, we estimate the surface gradients and the albedo using Least Squares (LS), Constrained Least Squares (CLS) and Total Least Squares (TLS) approach. We then obtain the high resolution depth map as well as image using Generalized Interpolation. These methods, being computationally less complex, can be used in real time applications. Results of experimentation on synthetic and real images are presented and the comparison of the three approaches is discussed.

1. Introduction

Spatial resolution refers to the number of pixels that are present in the image and is related to the level of detail that can be discerned from the image. High spatial resolution images of good quality are useful in many existing applications. For instance, in the fields of medical imaging, satellite imagery and in surveillance for security purposes, high resolution images are critical. Since, the resolution of an image is dependent on the device which is used to acquire the image, it is difficult to use very high resolution sensors as they are often expensive. Hence, there is a need to develop efficient methods to obtain good quality high resolution images given low resolution images.

3-D shape recovery of a scene is also used extensively for applications such as object tracking and recognition. There is a need of methods which upsample the image while preserving the 3-D structure. Most researchers use motion cue to increase the resolution of images by using super-resolution techniques where a number of low resolution observations of the same scene are used to increase the spatial resolution. Although the 3-D structure of the scene being imaged is inherently available from the disparity map, the motion cue being a 2-D feature matching technique does not consider the 3-D structure. The structure of the scene is encoded in the images in the form of shading, texture etc. If the image is upsamped and the high resolution depth of the scene is calculated from the high resolution image so obtained, the depth map will not be accurate as the shading, texture etc of the low resolution image may not be preserved in the upsampled image. Hence, techniques to obtain high resolution images which preserve the structure are required[1]. In [1], it has been shown that photometric cue with Generalized Interpolation can be used for estimation of high resolution depth map. It has also been shown in the paper that high resolution is possible if several low resolution images of the same scene, which are captured with different light source position, are available.

For real time vision applications, depth (surface gradients) estimation methods which are computationally efficient are required. However, many researchers use regularization based approaches in order to obtain better estimates. Now, if the cost used for obtaining the solution is non-convex, then optimization techniques such as simulated annealing are used to obtain the global minima, which makes these methods very time consuming. For instance, if we consider an assembly line where an object has to be moved from one place to another (industrial inspection), the requirement is to be able to calculate depth fast enough so that the assembly line functions smoothly. Here the requirement is speed not very high accuracy. In such a situation, the optimization methods are not useful.

In Literature, CLS and TLS have been used to for many applications. In [2], a CLS approach has been used to restore images blurred by a Gaussian impulse response where each point in the image has its own associated regularization parameter. An estimation method based on the CLS principle is proposed for the restoration of images distorted by a random point spread function and additive noise in [3]. In [4], Total Least Square method has been used to estimate the motion of the illuminant by observing an image sequence of an unknown object illuminated by it. A recursive total least square estimator is used to track the motion of the light source. In [5], TLS has been used for restoration of blurred and noisy images.
An algorithm for image restoration is proposed based on constrained total least-squares estimation, that is, adaptively regularized. Color Image restoration is done using TLS in [6]. Here, the blurring matrices as well as the observed image are contaminated by noise, therefore the TLS method is employed to restore the original image. In [7], blind channel equalization is done using TLS method.

In this paper, we present fast approaches for the calculation of high resolution depth map and intensity map using photometric cue. We use LS, CLS and TLS approaches for obtaining the upsampled images and depth map. Although the LS approach has been discussed in [1] and [8] for photometric stereo; we use this approach for comparison with other approaches. We show that the performance of these methods is comparable to regularization based methods. However, since the computational burden is low they can be applied to real time vision applications with acceptable accuracy.

2. Use of LS, CLS and TLS approaches with Photometric Stereo

In photometric stereo, the surface gradients of an object are estimated using a number of images taken under different light source positions. When the light source position is changed, the reflectance map of the object changes, but the gradients (being surface property) remain the same. In addition, different parts of the surface have different light reflectance property. If a Lambertian surface is assumed, the image irradiance equation relating the surface gradients and albedo can be written as:

$$E(x, y) = R(p(x, y), q(x, y)) = \rho(x, y)\hat{n}(x, y).\hat{s} \quad (1)$$

where $p(x, y), q(x, y)$ are the surface gradients in $(x, y)$ directions respectively. Here $\rho(x, y)$ represents the albedo, which is nothing but the fraction of light reflected from the surface at the point $(x, y)$ and its value lies between 0 and 1. $\hat{n}(x, y)$ denotes the surface normal given by $\frac{(-p(x, y), -q(x, y), 1)}{\sqrt{p(x, y)^2 + q(x, y)^2 + 1}}$, and $R(p(x, y), q(x, y))$ is the reflectance map. $E(x, y)$ is the image irradiance (or intensity) at point $(x, y)$ in the image. $\hat{s}$ is a unit vector in the direction of light source.

The surface gradient and albedo at a point are related to the intensity at that point according to Eq. (1). Since there are 3 unknowns $p(x, y), q(x, y), \rho(x, y)$, it is possible to obtain a unique solution using 3 linearly independent equations. We can obtain 3 equations for $E(x, y)$ from images captured using different light source positions. In real scenario, due to erroneous observations, the equations may be inconsistent and hence one needs to capture more than 3 images with different light source positions and obtain the solution using overdetermined set of equations using suitable error minimization technique such as the Least Squares or the Total Least Squares method.

2.1. Least Square approach

If we now consider $N$ images captured using $N$ different light source positions, we have $N$ equations. Putting them in matrix form using Eq. (1) we have,

$$\rho(x, y).\hat{n}(x, y) = S^{-1}.E \quad (2)$$

where $E$ is a $N \times 1$ matrix containing the image intensity values and $S$ is a $N \times 3$ consisting of the unit vector defining the light source direction. The LS solution is obtained by first finding $S^{-1}$ using Singular Value Decomposition (SVD) and then using Eq. (2) to solve for the surface gradients and albedo. For details refer to [8].

2.2. Constrained Least Square approach

Improvement over the LS method can be obtained by putting a constraint to the solution. One such approach is Constrained Least Squares (CLS) where Eq. (2) is solved using LS method and in addition, a constraint is put on the solution. Since the surface normal $\hat{n}(x, y)$ is always unity, it can be used as a constraint in the case of photometric stereo. In this case we want to find a vector $\rho(x, y)\hat{n}(x, y)$ such that $||E - S\rho(x, y)\hat{n}(x, y)||^2$ is minimized while maintaining $||\hat{n}(x, y)||^2 = 1$. This problem can be solved using Lagrange multipliers. The cost to be minimized is as follows,

$$\phi = ||E - S\rho(x, y)\hat{n}(x, y)||^2 + \lambda(||\hat{n}(x, y)||^2 - 1) \quad (3)$$

where $\lambda$ is the Lagrange multiplier. Putting a constraint on the solution restricts the solution to a proper subset of $\mathbb{R}^3$ (in this case). Thus, using the CLS better solution can be obtained since the solution space is now restricted as compared to the LS case. The solution obtained using CLS method depends on the value of $\lambda$, which can be estimated iteratively or one can use solution of a nonlinear equation to obtain it as a closed form solution.

2.3. Total Least Square approach

The TLS approach is another way to obtain improvement over the LS method. LS approach considers errors only in the intensity values. In real scenario, while using photometric stereo, observations pertaining to light source positions as well as the image intensity values can be noisy. So, a better solution may be obtained if errors in both source positions and intensity values are considered. The $E$ matrix in Eq. (2) contains the image intensity values which may be corrupted by sensor noise. Also, the $S$ matrix on the right hand side contains the estimated values of light source positions, hence it may be considered as a noisy measurement. So, Eq. (2) can be written as,

$$\rho(x, y).\hat{n}(x, y) = (S + \delta S)^{-1}(E + \delta E) \quad (4)$$
The solution is obtained by taking the Singular Value Decomposition(SVD) of the augmented data matrix \((S|E)\). See [9] for more details of TLS method. If the SVD of \((S|E)\) is given as,

\[
(S|E) = U \Sigma V^T
\]

where \(U\) and \(V\) are orthogonal matrices of dimensions \(N \times 4\) and \(4 \times 4\) respectively. \(\Sigma\) is a \(4 \times 4\) diagonal matrix.

Consider matrix \(V\) in partitioned form as follows

\[
V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}
\]

where \(V_{12}\) is a \(3 \times 1\) matrix and \(V_{22}\) is a scalar. The TLS solution at a particular pixel \((x, y)\) is given by

\[
\rho(x, y), \hat{n}(x, y) = -V_{12}V_{22}^{-1}
\]

3. **Generalized Interpolation and high resolution depth and intensity estimation**

Suppose we want to interpolate a function \(g(x)\). Consider the following abstract parametric decomposition of the function

\[
g(x) = f(a_1(x), a_2(x), \ldots, a_m(x)) \tag{6}
\]

where \(a_i(x), i = 1, 2, \ldots, m\) are different functions of the interpolating variable \(x\) and when they are combined by an appropriate \(m\)-variate function \(g\), one recovers the original function. If the functions \(a_i(x), i = 1, 2, \ldots, m\) are assumed to be continuous then we can interpolate the individual functions and then combine them using the function \(f\). We can obtain interpolated \(g(x)\). Such an interpolated technique is called generalized interpolation[1].

As mentioned above, Generalized interpolation can be used with the photometric stereo for image upssampling. Suppose we want to zoom an \(M \times N\) image and we are given a number of images of the same scene captured with different light source positions. As explained earlier, we can obtain albedo \((\rho(x, y))\) and surface gradients \((p(x, y), q(x, y))\) using photometric stereo and suitable minimization method such as LS, CLS or TLS. Now, the surface gradients and the albedo can be interpolated individually using a suitable interpolation technique. If \(k\) is the upsampling factor then each of the three spaces, \(p(x, y), q(x, y)\) and \(\rho(x, y)\) are of dimension \(kM \times kN\). We can, now, use the image irradiance equation to combine them to get the \(kM \times kN\) high resolution image, for each light source position.

The quality of the interpolated image and depth map depends on the values of albedo \((\rho(x, y))\) and surface gradients \((p(x, y), q(x, y))\) (since these are interpolated individually). Using CLS and TLS we get better estimates of surface gradients and albedo and hence it is possible to get a better reconstruction of the high resolution image and depth map when compared to the case when photometric stereo solution is obtained using LS approach.

The advantage of using Generalized Interpolation i.e. interpolating in the surface gradient and albedo domain lies in the fact that the change in image intensities may not always correspond to a change in depth. Factors such as position of illuminant, incident light intensity and reflectance properties of the object being observed also affect the image intensities. However, the variation in the shape of the object being imaged is generally smooth. So, interpolating the surface gradients and albedo gives better reconstruction of the high resolution depth map and the high resolution intensity map.

It may be mentioned here that one can use any one of the methods (LS, CLS or TLS) as initial estimates and obtain better estimates of high resolution fields using optimization techniques[8]. However, when the cost becomes non-convex, the computational complexity of the minimization algorithm is high and this limits the use of optimization methods in real time applications. Considering this fact, we show that the TLS and CLS approaches which are quite fast can be used for depth and intensity estimation with good accuracy. The idea of using TLS and CLS is to keep the computational cost burden low, so that real time processing is possible, which is very much necessary in practical vision applications.

4. **Discussion**

In this section, we present some of our experimental results to compare the performance of LS, CLS and TLS approaches. We have conducted experiments on both simulated and real images which were simulated or captured with different light source positions. For real images, the distance of the camera from the object is kept much larger than the object size so that orthographic projection can be assumed. Also, the light source is placed at a sufficiently large distance from the object, so the light source direction can be assumed to be constant for the whole surface. Note that the images presented in this section are best viewed on a high resolution computer screen.

To measure the performance of these methods quantitatively, we compare the mean square error between the recovered high resolution images by LS, CLS and TLS methods and the original image. We use Peak Signal to Noise Ratio (PSNR) as a figure of merit for comparison.

The first image that we consider is a synthetic image of a hemisphere that has a checkerboard pattern on it. In this case, the depth variation of the object is smooth whereas the intensity variations are abrupt. Eight images
of size $64 \times 64$ with different light source positions were generated through a computer program and LS, CLS and TLS were used to calculate $p, q, \rho$ fields with photometric stereo. Fig. 1A shows the actual high resolution image of size $128 \times 128$ with the source position $(0, 0, 1)$. Fig. 1B, 1C and 1D show the high resolution intensity maps with upsampling factor of $k = 2$ for the source position $(0, 0, 1)$ obtained using LS, CLS and TLS methods respectively. The value of $\lambda$ has been iteratively adjusted to 0.001 in the case of this image. A high value of $\lambda$ leads to a smooth solution and hence the discontinuities are not preserved. Small value of $\lambda$ gives a solution close to that obtained by the LS method, since the weighting given to the constraint is small. So, the value of $\lambda$ was set to an intermediate value in order to obtain a better solution using limited number of iterations.

The upsampled image recovered using TLS and CLS is better as compared to LS. See Fig. 1(B-D). The innermost circle of the checkerboard is not perfectly recovered with LS (shown in Fig. 1B) as compared to the high resolution images recovered with CLS and TLS approaches. For the Checkerboard image the true depth map is available, which is shown in Fig. 2A. The recovered high resolution depth with LS, CLS and TLS methods is displayed in Fig. 2(B-D). In the high resolution depth map of the Checkerboard image recovered using LS, the gradation of intensity values from the center of the hemisphere to the periphery is not proper. The upsampled depth maps recovered using CLS and TLS show a clear gradation of intensity from the center to the boundary of the hemisphere, indicating better depth estimation. The intensity rings on the depth maps are due to scaling done to display the recovered depth values as images. The performance of TLS as well as CLS approach is better in terms of the PSNR (Table I) when compared to LS. The performance of CLS approach depends on the value of $\lambda$ and hence may perform better or worse than the TLS approach. In the case of TLS approach, the performance depends on the threshold put to reject small values of the scalar $V_{22}$.

Next we consider a white soft toy Jodu which when imaged gives smooth variations in intensity but has arbitrary depth variations. Eight images of Jodu were captured with different light source positions. Fig. 3A shows the captured image of the soft toy of size $234 \times 234$ with the light source position $(-0.838, -0.719, 1)$. In order to compare the performance of the three methods LS, CLS and TLS with photometric stereo, we have considered these eight images as the actual high resolution images. For our experiment, we downsampled these eight images by a factor of $k = 2$ so that we have eight images of size $117 \times 117$, which have been considered as observed low resolution images. We then obtain the low resolution $p, q, \rho$ fields using LS, CLS and TLS approaches from these images and use Generalized Interpolation to obtain high resolution intensity map and depth map.

Fig. 3B, 3C and 3D show the high resolution intensity maps of the Jodu image obtained using LS, CLS and TLS methods respectively. In Fig. 3A, we show the original image. The high resolution image obtained by using TLS reveals better details such as the discontinuities on the nose and fur on the body (See Fig. 3D). The shadow on the tongue is better revealed when TLS method is used as compared to the shadow estimated by LS and CLS methods. The performance of CLS method is better than LS due to the additional constraint on the solution, depending on the value of the lagrange multiplier $\lambda$. We have iteratively adjusted the value of $\lambda$ as 0.009. It may be mentioned here that the time required to arrive at a suitable value of $\lambda$ is much less when compared to the optimization methods, where one needs to adjust many parameters iteratively. Table I shows the PSNR values for the upsampled intensity maps. For the Jodu image, the performance of TLS is much better than LS and CLS.

In the case of the Jodu image, we do not have the true depth map since the laser scanner does not work well with such a furry object with discontinuities. Hence, for our experiments we consider the depth map retrieved from the high resolution images as the actual depth map. The surface gradients and the albedo are recovered from the undecimated observations (here $234 \times 234$ images) using photometric stereo and have been used as the true depth map or the ground truth to compare the performance of the LS, CLS and TLS methods. Fig. 4 gives the depth map of the Jodu image estimated using the three approaches. The three depth maps displayed using LS, CLS and TLS do not look much different, but on comparing the PSNR obtained for the depth field using LS, CLS and TLS we observe that the performance of TLS is better than the LS method and comparable to the CLS method.

We refer to [8] to compare the performance of our faster approaches with regularization based optimization technique. In [8], the surface gradients, albedo and the super-resolved image are modelled as separate Markov Random fields and a regularization scheme is used to super-resolve these fields. MRF based approach uses LS solution as an initial estimate. The performance of the MRF based approach depends on the values of the various parameters in the cost function. Since, there are several parameters to be iteratively adjusted, obtaining a better solution is very time consuming. The PSNR for depth values of the Jodu image obtained using LS (36.16 dB), CLS (45.84 dB) and TLS (44.66 dB) is comparable that of the MRF approach (41.82 dB) given in [8]. Hence, using LS, CLS and TLS methods we can obtain a solution faster and with accuracy comparable to that of regularization based optimization techniques. This makes these three methods a better choice for real time vision applications. Note that we cannot compare the PSNR for the intensity maps obtained with our approaches and MRF.
Figure 1: (A) Original Image of size $128 \times 128$ and source position $(0,0,1)$ and Estimated High Resolution Images for same source position with (B) LS (C) CLS (D) TLS methods

Based approach as the albedo of the images may be scaled differently in the two cases for display.

Figure 2: (A) Ground Truth and Recovered High Resolution Depth map for checkerboard image with (B) LS (C) CLS (D) TLS methods

Table 2 tabulates the processor time taken for low resolution surface gradients and albedo calculation using LS, CLS and TLS methods. All the simulations were done on a 1.3 GHz processor. The processor time taken by each of the three methods clearly depends on the size of the image. The LS method is the fastest in terms of the processor time. The TLS method takes more processor time than LS but better results are achieved by using TLS. For the CLS method, the major task is to adjust the lagrange multiplier, the value of which can be computed by using a non-linear equation solver such as the Newton’s method such that the constraint is satisfied. It can also be iteratively adjusted. Therefore, the CLS method is time consuming as compared to LS and TLS. Real time applications require a rate 40 ms per frame. It is evident from Table 2 that LS method satisfies this requirement. Using a higher speed processor, CLS and TLS would easily satisfy this requirement. As compared to these methods which produce the result is a few milli seconds, the optimization techniques like simulated annealing take hours for convergence.

Figure 3: (A) Original Image of size $234 \times 234$ and source position $(-0.8389,-0.7193,1)$ and Estimated High Resolution Images for same source position with (B) LS (C) CLS (D) TLS methods

5. Conclusion

In this paper, we have compared the LS, CLS and TLS approaches for photometric stereo for obtaining high resolution images and high resolution depth map. The performances of CLS and TLS methods are comparable and both perform better than the LS approach. We have also shown that the performance of these approaches is comparable to the regularization based approaches which are computationally very intensive. Though, in case of CLS, the value of $\lambda$ has to be adjusted iteratively, it is not really time consuming. Regularization based optimization techniques may perform better than LS, CLS and TLS approaches, but they are computationally very taxing and slow since the solution is obtained iteratively. The solution also depends on the particular type of regularization term used in the configuration while solving the problem.
Figure 4: Recovered High Resolution Depth map for dog image with (B) LS (C) CLS (D) TLS methods

Table 1: PSNR Comparison for Jodu image and Checkerboard image for a magnification factor of 2 for different source positions. The (DEPTH) row in the table gives the PSNR comparison for the depth field.

<table>
<thead>
<tr>
<th>Source position for Checkerboard image</th>
<th>Source position for Jodu Image</th>
<th>PSNR in dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0, 0, 1)</td>
<td></td>
<td>20.86</td>
</tr>
<tr>
<td>(DEPTH)</td>
<td></td>
<td>12.63</td>
</tr>
<tr>
<td>(0.8389, 0.7193, 1)</td>
<td></td>
<td>22.69</td>
</tr>
<tr>
<td>(-0.1763, -0.5596, 1)</td>
<td></td>
<td>19.03</td>
</tr>
<tr>
<td>(DEPTH)</td>
<td></td>
<td>36.16</td>
</tr>
</tbody>
</table>

Table 2: Processor time for low resolution surface gradients and albedo calculation using LS, CLS and TLS methods

<table>
<thead>
<tr>
<th></th>
<th>Processor time in seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LS method</td>
</tr>
<tr>
<td>For Jodu Image</td>
<td>0.062</td>
</tr>
<tr>
<td>For Checkerboard Image</td>
<td>0.015</td>
</tr>
</tbody>
</table>

* This does not include the time required for adjusting the parameter $\lambda$.

This restricts the use of such optimization techniques in real time applications. The experimental results show that the methods discussed here can be used in real time applications where speed not high accuracy is a priority.

6. References


