Detailed 3D reconstruction from multiview images

Jie Tan Siddharth Choudhary Parikshit Ram

1. The problem

While sparse 3D reconstructions of scenes suffice for popular applications such as photo-tourism, detailed 3D digital reconstruction of objects has applications in many fields such as computer graphics and cultural heritage preservation. Although laser scanning is usually employed for this purpose, detailed 3D reconstruction from multiview images is more cost-efficient. In this paper, we propose a method for detailed 3D reconstruction by combining techniques from multiview stereo and photometric stereo.

2. Related work

Many solutions have been proposed for the long standing problem of multiview stereo [7]. However, conventional multiview stereo algorithms perform poorly on texture-less objects. Cues from object silhouettes can be used for reconstruction [4]. Multiview stereo algorithms have also been extended to internet scale image collections with promising results [3]. However, these algorithms still lack detailed reconstruction in textureless regions.

Photometric stereo uses shading cues for very detailed 3D reconstructions of texture-less objects [2] and has been coupled with other visual cues for detailed multiview stereo. Hernandez et al. [5] use photometric information and the visual hull to reconstruct closed surfaces. The parallax and shading information have been combined to obtain high quality continuous depth maps [1]. Recently, a piecewise formulation to combine structure-from-motion and photometric stereo was proposed as well [6]. In contrast to these approaches, we will use coarse feature correspondences and photometric information to perform detailed reconstruction.

3. Approach

The main challenge of reducing multiview stereo photometric stereo is building pixel level correspondence. We will estimate the pixel correspondence from the coarse feature correspondence. Given a pair of images of the same object taken at different view point and lighting, we will use SIFT descriptors to detect features on each of the images

![Image 1](a) Bunny-1 (b) Bunny-2 (c) Bunny-3

![Image 2](d) Box-1 (e) Box-2 (f) Box-3

Figure 2. Images of the same object from different camera locations and lighting conditions – the first row presents the Bunny images and the second row presents the Box images. and subsequently use RANSAC to find correspondence of these feature points across two images.

We will associate a coordinate frame with each feature in the feature correspondences according to its location \((x, y)\), orientation \(\theta\) and scales \((s_1, t_1)\) as shown in Figure 1. Let a feature \(k\) be described by \((x_1, y_1, \theta_1, s_1, t_1)\) in image 1 and by \((x_2, y_2, \theta_2, s_2, t_2)\) in image 2. Then the affine transformation \(H\) of a feature in image 1 to its corresponding feature point in image 2 is given by \(H = T_2T_1^{-1}\) where

\[
T_t = \begin{pmatrix}
  s_1 \cos \theta_1 & t_1 \sin \theta_1 & x_i \\
  -s_1 \sin \theta_1 & t_1 \cos \theta_1 & y_i \\
  0 & 0 & 1
\end{pmatrix},
\]

With these coarsely sampled transformations \(H_i\) at each feature point \(i\), we will build up a smooth transformation field throughout the whole image using radial basis functions. For any pixel location \((x_1, y_1)\) in image 1, its corresponding location \((x_2, y_2)\) in image 2 is estimated as

\[
\begin{pmatrix}
  x_2 \\
  y_2 \\
  1
\end{pmatrix} = \left( \sum_i w_i H_i \right) \begin{pmatrix}
  x_1 \\
  y_1 \\
  1
\end{pmatrix},
\]

where the weights \(w_i\) for each \(H_i\) is determined as follows: for a pixel \(p\), only its \(k\)-nearest SIFT features \(\{i_1, \ldots, i_k\}\) get non-zero weights \(w_{ij} = 1/d(p, i_j)Z\) where \(d(p, i)\) is the squared Euclidean distance between pixel \(p\) and feature \(i\) and \(Z = \sum_{j=1}^{k} 1/d(p, i_j)\) is the normalization factor.

Thus, the images are transformed to align on pixel level. At this point, we use a photometric stereo (PS) algorithm [2, Algorithm 2.2] to find out the color map, the normal field and the depth map which can be used for reconstructing a detailed 3D shape. The details of the PS algorithm are omitted here for the lack of space.
4. Evaluation

For our evaluation, we generated a ground-truth dataset by creating a virtual 3D scene in Maya CAD software and setting up several different camera locations and lighting directions. We created two sets of three images -- the Bunny and the Box images shown in Figure 2. Each image of the same object is taken from a different camera locations and lighting conditions. Using the first image as the base image, we set about the task to transform the second and third images to align with the first image. We begin by obtaining the SIFT feature correspondences using RANSAC for each pair of images (between images 1 and 2 and between images 1 and 3). The matched SIFT features are presented in Figure 3. For the highly textured Bunny images, RANSAC is able to obtain a large number of accurate correspondences. However, for the Box images, RANSAC obtains correspondences on the textured part of the box, but finds no correspondences in large parts of the box.

Given these matched SIFT features, we obtain the affine transformation for each feature and use the 10-nearest SIFT features of each pixel to build the pixel level correspondences. For each set of images, the second and third images is transformed to the coordinate axis of the first image. The resultant images are presented in Figure 4. The alignments of second and third Bunny images are fairly indistinguishable from the first image. The Box image alignments are not as precise and a lot of straight lines are broken and some parallel lines in the image are not parallel anymore.

With these 3 images effectively taken from the same camera location and different lighting conditions, we perform photometric stereo. For evaluation purposes, we obtain the ground truth results of PS of the object with the camera location of the first image by using multiple images from the first camera location and different camera lightings. We also present the results of PS on the input images without any transformation to validate our hypothesis that our proposed locally-affine smooth transformations benefit PS. The results are presented in Figures 5 & 6. Table 1 presents the relative error of the color and depth maps obtained by our proposed method. The relative error corresponds to the ratio of pixel-wise sum-of-squared-differences (SSD) of the result of our method from the ground truth and the SSD of the result of PS on the input images from the ground truth. Figure 7 presents the distribution of the pixel-wise error (in terms of angle) of the normal field.

<table>
<thead>
<tr>
<th>Images</th>
<th>Color map</th>
<th>Depth map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bunny</td>
<td>0.4916</td>
<td>0.1650</td>
</tr>
<tr>
<td>Box</td>
<td>0.3968</td>
<td>0.2675</td>
</tr>
</tbody>
</table>

Table 1. Relative error of the color and depth maps obtained from the aligned images with respect to the color and depth maps obtained directly from the input images.
Figure 6. Results of PS on the Box images – the notations are same as in Figure 5.

Figure 7. Distribution of the error in the normal fields – the notations are same as in Figure 5.

For both the Bunny and Box images, the results of PS applied to the input images (the center column of Figures 5 & 6) are fairly different from the ground truth color and depth maps and the normal field (the left column of Figures 5 & 6). The results of PS on the transformed (aligned) images (the right column of Figures 5 & 6) are closer to the ground truth, being visually indistinguishable in case of the Bunny images and slightly less so for the Box images. Table 1 indicates that the alignment removes more than 50% and 60% of the error in the color map, and around 80% and 75% of the error in the depth map for the Bunny and Box images respectively. The normal field error distributions in Figure 7 indicate that the error in the normal field also reduces significantly with the transformed images.

5. Discussion

The results indicate that our proposed method of locally-affine transformations for image alignments significantly improve the quality of PS. As expected, the quality of PS on the aligned images depends on the quality of the alignments. The alignments for the Bunny images are fairly indistinguishable to the base image (Figure 1 (b-c)) and so are the results of PS (Figures 5 & 7(a)). The results for the Box images in Figure 6 are visibly less accurate. This is probably because of the improper alignment of the images which deformed the object (Figure 1(f)). We list the probable causes for improper alignments:

- The lack of feature matches at sharp corners in the image can hurt the alignment (and hence the reconstruction). This was the possible reason for the breaking of the edges of the box in Figure 1(e-f). For simple objects, it is possible to manually match sharp corners to assist the alignment. For more complex objects, the hope is that the sharp corners are unique enough to be matched automatically by RANSAC.

- This brings us to another reason for improper alignment – wrong feature matches by RANSAC. Our alignment scheme is sensitive to wrong matches and a single wrong match can significantly hurt the transformation. One way to avoid this issue is by manually removing wrong matches. A more automatic scheme to remove wrong matches can employ three images for the matching – if a feature i in image 1 is mapped to a feature j in image 2 and feature k in image 3, then discard this feature match unless feature j in image 2 matches with feature k in image 3.

- Another reason for improper alignment is the fact that we use the Euclidean distance on the image plane for finding the k-nearest features for a pixel. This can severely deform areas around sharp corners. Using the actual world distance would avoid this problem but the actual world distance is not easily accessible. One computationally expensive way is to create a coarse 3D reconstruction using structure-from-motion techniques and use the distance along this coarse 3D surface.

Other possible avenues for future work include (i) devising a smarter way of choosing weights \( w_i \) in Equation (2) for the locally-affine transformations, (ii) incorporating more images (taken from different camera locations and lighting conditions) to boost the quality of the results, and (iii) applying our proposed method to more noisy internet images.

References