Automated 3D Reconstruction

Frank Wang (In collaboration with Ravi Netravali)

Abstract

In this report, we present a 3D reconstruction method that uses two uncalibrated images to construct the final 3D object. It uses both SIFT and Harris corners to provide a more representative modelling of the object. We found SIFT to miss important features, which can be rectified using Harris corners. We use these feature points to calculate the fundamental matrix. Moreover, rather than calibrating each image individually, we choose to calibrate the camera using Zhang's method so that the calibration only has to happen once. We show results from our experiments to show the effectiveness of this method.

1. Introduction

Reconstructing a 3D object from input images is an essential task in the field of computer vision. Recent advances in photo databases have led to an increase in the use of 3D reconstruction. While there have been many recent advances in 3D reconstruction, most new techniques still suffer from the same problems which plagued original methods.

First, many approaches require a camera calibration for each image used in the 3D reconstruction. This calibration is necessary to obtain the intrinsic and extrinsic parameters of the camera which are used to project points from 2D images into 3D space. Performing calibrations for each image is impractical for new approaches that emphasize using as many images as possible for the reconstruction.

The other setback relates to user input. Many approaches require users to pick points on the input images that specify either the object to reconstruct or the planes used to perform the reconstruction. While user input may improve the reconstruction as it improves depth detection and crops the image so only necessary points are used, it does not encourage using many input images. Several techniques overcome these setbacks by performing reconstruction based on uncalibrated images obtained by the same, freely moving camera. Camera calibration are performed once and the resulting parameters are used for the reconstruction across all input images. Capturing pairs of feature points from the input images is an important step in the 3D reconstruction process. There are several feature point detection algorithms which are currently used by many reconstruction techniques, though each has its own setbacks. Two popular approaches are Scale Invariant Feature Transform (SIFT) and corner extraction algorithms such as Harris corner detection. Each approach yields pairs of matched feature points across the input images.

In this project, we develop a new technique which reconstructs a 3D object from two input images captured by a freely moving camera. We then combine SIFT and Harris corner detection to obtain the relevant pairs of feature points from the two input images. Prior to using these matching points, we use the Random Sample Consensus (RANSAC) algorithm to remove outlier matches. We calculate the intrinsic parameters of the camera used to take the input images with Zhangs method. Using the intrinsic parameters of the camera and the matched feature points from the input images, we are able to obtain a 3D reconstruction for the primary object in the image.

The report is organized as follows. Section 2 describes related work, and then section 3 provides an overview of our algorithm. Then, the later sections delve into more details. Section 4 and 5 talk about camera calibration and automated feature detection, which are tools that we use in our reconstruction. Section 6 talks about the epipolar geometry, and section 7 talks how to use the epipolar geometry to do the 3D reconstruction. We show some results in section 8, and we conclude in section 9.

2. Related Work

There has been much work on reconstructing 3D object from multiple images using epipolar geometry. A good review of the methods and previous work in this area can be found here [**?**]. Our project uses many of these techniques and combines them with more modern automation techniques, such as SIFT and Harris corners to reduce the amount of human interaction required for the reconstruction. Our project is primarily based on the work of Peng et al. [**?**], where they use these automated techniques to perform the 3D reconstruction.

There is also a separate set of work that tries to do 3D reconstruction using only a single uncalibrated image [**?**, **?**]. However, these type of reconstructions require some

prior knowledge about the parameters of the scene and the camera. Multiple view 3D reconstruction does not require any prior knowledge because we can extract the necessary parameters from the images.

3. Overview

In this section, we give an overview of our algorithm, which we will describe in more detail in later sections.

1. We find the intrinsic parameters of our camera using Zhang's method.
2. Using SIFT and Harris corner detection, we identify pairs of features points on both images, and we use RANSAC to find the inliers so that we can generate a representative fundamental matrix.
3. With the intrinsic parameters of our camera and the fundamental matrix, we can derive the essential matrix.
4. The essential matrix is used to derive the projection matrices of the images which each comprise of a rotation matrix and translation vector. With this information, we can reconstruct the 3D image using triangulation.
5. With this information, we can reconstruct the 3D image using triangulation.

4. Camera Calibration

To calibrate our camera and obtain its intrinsic parameters, we use Zhangs method. Zhangs method is a technique that uses several images to derive a cameras focal length, aspect ratio, and principal points. While there are many other approaches to performing this essential task, Zhangs method is more flexible and robust. The two traditional approaches are often categorized as either photogrammetric or self-calibration. Photogrammetric methods require using a 3D object whose 3D coordinates are precisely known. Taking multiple images of this 3D object lets one infer the intrinsic parameters of the camera as they can be derived from the difference between the actual coordinates of the object and what is seen across the images. However, the apparatus necessary to perform this type of calibration is expensive. Self-calibration, while less costly, is not reliable as there are often not enough known points to estimate all the necessary parameters. Self-calibration requires moving the camerain a static setup and performing a feature points matching across the taken images. Self-calibration also uses constraints pertaining to the rigidity of the object considered.

Zhangs method is a cross between photogrammetric and self-calibration techniques. Zhangs method requires one to construct a pattern on a paper and attach it to a planar surface. Several images of the pattern, from different angles, are then taken. As long as the camera or the pattern is stationary, the movement is not restricted. Using an understanding of the geometry of the designed pattern, constraints on the intrinsic parameters arise from each view. Using all of these constraints, a set of intrinsic parameters which satisfy all the considered views can be inferred. Of course, using more views will yield intrinsic parameters closer to their actual values. Thus, Zhangs method incorporates understanding 3D coordinates of the pattern (photogrammetric) and using multiple views to set constraints on the intrinsic parameters (self-calibration), but it only considers one plane and a user-designed pattern. In their paper, Zhang et. al. show that their calibration technique is both flexible, robust, and accurate.

A cameras intrinsic matrix is defined as:

$$A = \begin{bmatrix} \alpha & \gamma & u0 \\ 0 & \beta & v0 \\ 0 & 0 & 1 \end{bmatrix}$$

where (u0, v0) is the principal point, α and β are image scale factors, and γ is a parameter which describes the skew in the image axes. We use the pinhole camera model which dictates that a 3D point, M, is related to its image projection, m, by:

$$sm = A \begin{bmatrix} R & t \end{bmatrix} M \tag{1}$$

where R is the rotation matrix and t is the translation vector of the considered image. For each view used, we estimate a homography matrix:

$$H = \begin{bmatrix} h_1 & h_2 & h_3 \end{bmatrix} = \lambda A \begin{bmatrix} r_1 & r_2 & t \end{bmatrix} \tag{2}$$

where r_1 and r_2 are the components of the rotation matrix, R, and lambda is a scaling factor. Since r_1 and r_2 are orthonormal, we can derive the following two constraints:

$$h_1^T A^{-T} A^{-1} h_2 = 0 \tag{3}$$

$$h_1^T A^{-T} A^{-1} h_1 = h_2^T A^{-T} A^{-1} h_2 \tag{4}$$

The approximation of the cameras intrinsic parameters is improved using a maximum-likelihood estimation. Zhang's method also extracts the cameras extrinsic parameters for the given input images, but we do not make use of this data as we are reconstructing an object separate from the considered planar pattern.

5. SIFT and Harris Corners

Here, we describe two forms of automated feature detection. First, we discuss the scale-invariant feature transform (SIFT) [**?**], which is an algorithm that detects local features in an image. For our purposes, we use SIFT to gather matching points between the two images. However, many

times SIFT is insufficient to capture important features specific to the images. In our case, we found it difficult for SIFT to capture the corners, which define the shape of the object. As a result, we also applied Harris corners [**?**] and included them as features. The difference in the reconstructions with and without Harris corners are show in Section 8.

SIFT is an algorithm that detects local features in images. It is widely used because of its robustness to geometrical changes as well as its ability to successfully extract stable feature points. Since we just use SIFT as a black box package, we will only provide a high level overview. Here are the four main steps in the SIFT algorithm: extreme point detection, accurate localization of key point, assignment the main orientation of key point, and the creation of key point descriptor.

To detect the extreme point, a Difference of Gaussian (DOG) pyramid of the image is built, and a pixel is compared with its 26 neighbors in 3x3 regions. If the point is a maximum or minimum in the 26 neighbors in the DOG scale space, then it is a feature point. In order to assign the orientation of the key point, samples around the key point are taken. A gradient histogram is created from the gradient orientations of the sample points. From the histogram, we assign the highest peak as the orientation of the key point. Finally, the key point descriptor is created by sampling points with a 16x16 region around the key point and creating 8 orientation bins from each of the 4x4 subsections. From this, we can create seed points to give us a feature vector. SIFT decides that two key points are matching by checking the Euclidean distance between two feature vectors as well as the nearest neighbor algorithm. More specific details can be found in the paper [**?**].

We also use Harris corner detection as a means to improve our 3D reconstruction. A corner is defined as a point in the image where the gray changes drastically or the junction of the contour boundary.

The Harris corner detection technique is an enhancement of Moravec corner detection. For each pixel in a considered image, a Moravec corner detector creates a patch centered at that pixel and compares it to neighboring patches which have large overlap. Neighboring patches are created at 45 degree shifts (perpendicular, parallal, diagonals). For each comparison, the algorithm computes the sum of squared differences between the two considered patches with respect to the difference in gray scale intensity. The algorithm is able to deduce what kind of feature point the pixel represents simply from several squared difference sums, which represent different orientations of compared patches. If the considered pixel is on a edge, the compared patches will differ from the considered patch significantly as we move perpendicular to the edge, and will be similar to the considered patch as we move parallel to the edge. Similarly, a pixel representing a corner will yield patches which significantly differ from the considered patch. However, the considered patch will differ greatly with all considered patches, regardless of orientation. If the considered pixel is not on an edge or a corner, the squared differences obtained will be relatively small regardless of orientation.

Harris corner detection improves on Moravec corner detection in several ways. Harris corner detection considers shifts for the neighboring patches in all directions rather than just 45 degree shifts. Additionally, Harris corner detection uses a circular Guassian window for comparisons which reduces noise. As a result, Harris corner detectors are able to distinguish between edges and corners more accurately than Moravec corner detectors.

These matching points are important to calculate the 3D geometry of our object.

6. Geometry of 3D Reconstruction

In order to understand the purpose of the fundamental and essential matrix, we first describe epipolar geometry. Epipolar geometry captures the intrinsic projective geometry between two cameras views, which only depends on the internal parameters of the camera and relative positions of the two views. The fundamental matrix contains this information, and from the fundamental matrix, we can derive the essential matrix, which gives us the relative position and orientation of the two camera views. From that, we can derive the actual 3d points of the points in both views.

6.1. Epipolar Geometry

In our project, we take two uncalibrated pictures of the same object but from different views. Epipolar geometry gives us a relationship between these images.

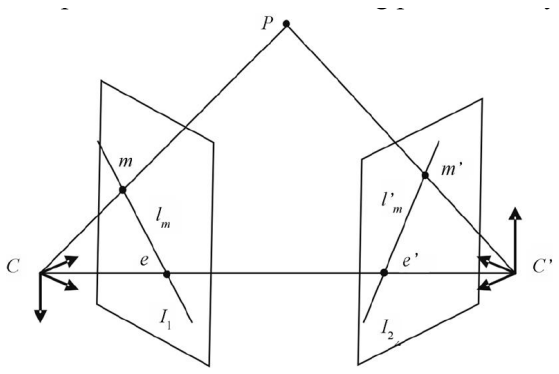

Figure 1. Epipolar geometry of two images from different views.

Figure 1 illustrates the relevant epipolar geometry. I_1 and I_2 are the images planes and C and C' represent the optical center of the camera from different views. m and m' represent P projected onto I_1 and I_2 respectively. e and e' represent the epipoles, which are the intersections of the

image planes with the line connects the camera centers. l_m and l'_m are the epipolar lines, which is where the epipolar plane intersects the image plane. It is important to note that m and m' lie on the l_m and l'_m respectively. As a result, we know that the matching point for m is going to be on the line l'_m in I_2 instead of anywhere in space of I_2. In order to capture this geometric restriction, we use the fundamental matrix, which will describe in the next section.

6.2. Fundamental Matrix

The fundamental matrix F is a mathematical representation of the epipolar geometry described above. From Figure 1, we can see that for a point m in I_1, there is a corresponding epipolar line l'_m in I_2 and the matching point m' must lie on that line. We can think of this as a mapping from a point to a line, specifically a projective mapping from points to lines. This projection is represented by the fundamental F. Algebraically, two matching points m and m' on two images must satisfy the following relation:

$$m^T F m = 0 \tag{5}$$

There are many methods to find the fundamental matrix for a specific pair of images algebraically and geometrically. Each method has its tradeoffs for time, complexity, and error. For the sake of space, we provide a high level overview of how we calculated the fundamental matrix as many of the details vary based on implementation. The fundamental matrix is a 3x3 matrix, so there are 9 parameters. However, after normalizing on one parameter, we only need to find 8 parameters. This means we need at least 8 pairs of points to construct the matrix. To obtain points for corresponding features in the two images automatically, we used SIFT and Harris corners as described in Section 5, but there are also many other ways to obtains these points. However, in most of these methods, we have many more than 8 points. We use RANSAC [**?**] to filter out outliers and calculate F such that there are the largest number of inliers. To calculate the fundamental matrix, we use the least median squares method [**?**] to minimize error.

We now want to obtain the extrinsic parameters (orientation and translation) from the fundamental matrix, which are found in the essential matrix.

6.3. Essential Matrix

The essential matrix gives us the extrinsic parameters of the two views. In other words, we can find the translation and rotation of the views relative to each other. However, we can only use the essential matrix if we know the internal parameters of the camera. In section 4, we describe a method to find the internal parameters of the camera. With that, we have the following relation between essential matrix and fundamental matrix:

$$E = K^T F K \tag{6}$$

F is the fundamental matrix and K is the internal parameters of the camera.

Theoretically, the essential matrix should have two equal non-zero eigenvalues and a zero eigenvalue. However, due to the noise in the data, this usually is not true in practice. In order to rectify this, we use Singular Value Decomposition (SVD) [**?**] to capture the diagonal matrix with the eigenvalues. We set the smallest eigenvalue to 0 and set the other two eigenvalues as the average of each other. Then, with this diagonal matrix, we construct a new essential matrix.

Now, we have all the tools to perform the 3D reconstruction.

7. 3D Reconstruction

Having all the necessary tools, we now reconstruct the 3D image. First, we describe the perspective project transform from 3D to 2D using the projective matrix P. It is important to note that we use the pinhole camera model. Let m be the 2D image point in homogeneous coordinates and M be the point in 3D space in homogeneous coordinates. We use the projective matrix P to related m and M.

$$m = PM = K[R|t]M \tag{7}$$

K is the intrinsic parameters of the camera, R is the rotation matrix, and t is the translation matrix.

We now need to create the projective matrices for each image. For simplicity, we set the first image to be our world coordinates, so our projective matrix is the following:

$$P_1 = K[I|0] = [K|0] \tag{8}$$

Now, we need to find the projective matrix for the second image P_2 relative to the world coordinates represented by P_1. More specifically, we need to find the rotation and translation matrix. Since we calibrated the camera using Zhang's method, this can be done using the essential matrix. We take the SVD of the essential matrix E to decompose it and we suppose the following W:

$$E = USV^T, \text{ suppose } W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

There are two possible values for the rotation matrix: $R = UWV^T$ or $R = UW^TV^T$. There are also two possible values for the translation vector: $t = u_3$ or $t = -u_3$ where u_3 is the last column of U. Since $P_2 = K[R|t]$, we have four possible values for P_2. In order to pick the correct one, we calculate the 3D spatial coordinate for one pair of matching features. We then pick the P_2 such that the point is in front of both cameras. In our case, this means that the Z coordinate of the 3D coordinate is positive.

After we have the projective matrix for both images, we calculate the 3D spatial point $M = (X, Y, Z, 1)$ for each set

of matching points $m = (u_1, v_1, 1)$ and $m' = (u_2, v_2, 1)$. Suppose P_{1i} and P_{2i} are the ith row vector of P_1 and P_2 respectively. To obtain the 3D spatial coordinate M, we have the following equation to do triangulation for each m, m' pair:

$$\begin{bmatrix} P_{13}u_1 - P_{11} \\ P_{13}v_1 - P_{12} \\ P_{23}u_2 - P_{21} \\ P_{23}v_2 - P_{22} \end{bmatrix} M = 0$$

We can use the least square method to find M for each pair of matching points. Once we have all the 3D spatial points, we can reconstruct the image.

8. Results

Here, we show some of our reconstructions as well as illustrate the implementation of our immediate steps, such as the SIFT matching points and Zhang's method. We We used a Panasonic Lumix DMC-ZS5 camera to take pictures.

8.1. Camera Calibration

As stated above, we use Zhang's method to calibrate our camera. We use a sheet of paper with six 2" by 2" black boxes. Some pictures of the different angles used for the calibration are shown in Figure 2.

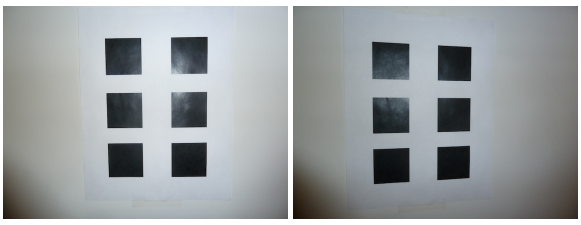

Figure 2. Different calibration angles for Zhang's method

We used multiple views and took the average over different permutations. We found the following intrinsic parameter matrix K for this camera:

$$K = \begin{bmatrix} 360.506 & -38.7974 & 128.692 \\ 0 & 470.953 & 162.864 \\ 0 & 0 & 1 \end{bmatrix}$$

8.2. Reconstruction of Zhang's images

We first reconstructed the images provided by Zhang [**?**] because we were provided with the pictures and internal parameters. This provides us with proof that our algorithm works with another camera other than our own. Similarly, we wanted to a baseline reconstruction to ensure that our camera calibration was mostly accurate. Figure 3 show two original images from different views used for the reconstruction. Figure 4 show views of the 3D reconstructed object without including the corners as features. We can see that the normals of the different planes are pretty well represented as the normals appear to be almost perpendicular to each other. Similarly, the edge of the box is also pretty well reflected in our reconstruction. However, we did not include the corners, which were not detected by SIFT, as a feature, so the box does not have the full rectangular shape. Figure 5 shows an improvement in the shape of the box when we included the corners as a feature.

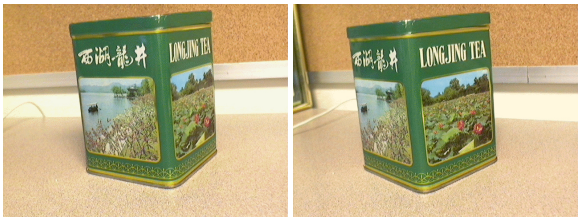

Figure 3. Original Zhang photos

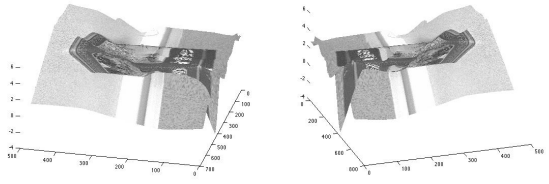

Figure 4. 3D reconstruction without corners included as features.

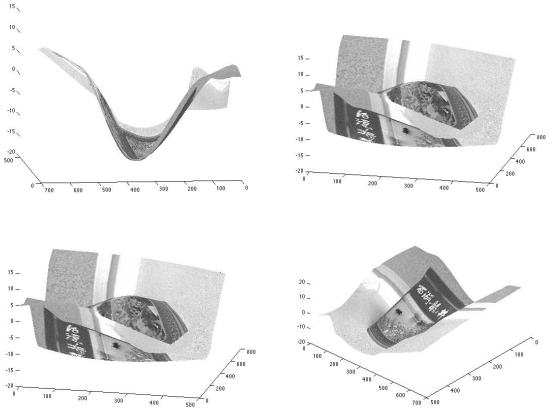

Figure 5. Different views of 3D reconstructions with corners included as features.

8.3. Our Reconstructions

These are the reconstructions that we did with our own camera. Figure 6 has the original images, and Figure 7 shows the matching points found by SIFT. Figure 8 shows our reconstruction without including the corners as features.

As we can see, like in the other reconstruction, the reconstruction is pretty representative of the object, but the shape is not very well preserved. However, with the addition of corners, the shape of the box becomes more representative of the actual object as seen in Figure 9. It is important to note that the reconstruction is not perfect because we only used two images, so we could only gather a limited number of information and constraints about the spatial points. This could be improved with more images and a more refined fundamental and essential matrix.

Figure 6. Original Images

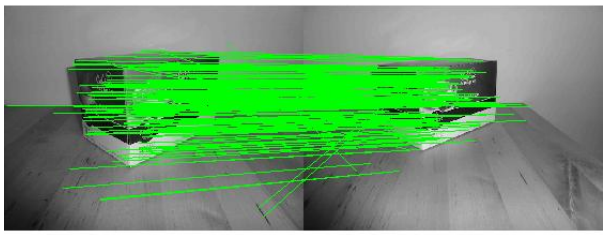

Figure 7. Matching points on images.

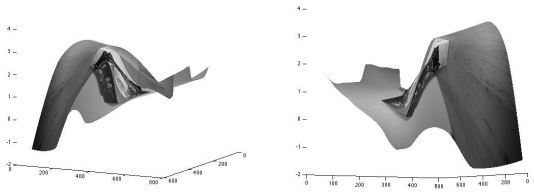

Figure 8. 3D reconstruction without corners as features.

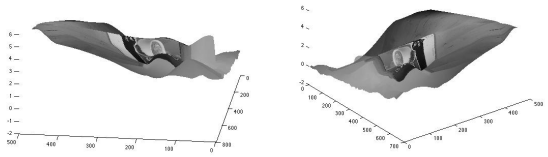

Figure 9. 3D reconstruction with corners as features.

9. Conclusion

This project describes a flexible and practical 3D-reconstruction technique. Rather than having to perform a camera calibration for each input image used in the reconstruction, as many previous techniques have required, we only perform the calibration once using Zhangs method. Thus, this approach scales well with the number of input images used, as long as they are all taken with the same camera. We also combine several common feature point extraction methods to enhance the detail in the final 3D reconstruction. Using both SIFT and Harris corner detection, we are able to get enough pairs of matching points to recover all eight unknowns in the fundamental matrix while incorporating all the necessary feature points to capture depth and structural details in the reconstructed object. The benefits of including Harris corner detection are highlighted in the results we provided. Both edges and corners are much more pronounced in the reconstructed images using corner detection than in those when we used SIFT alone.