# Project Final Report

## Stereo Reconstruction - group 10

**Authors: Barry (Shichen) Hu, Jiongyan Zhang, Ran Ding**

## 1 Introduction

Stereo Reconstruction is a problem that has occupied computer vision researchers for the past few decades. In this project, we aim to build a complete pipeline for stereo reconstruction, which is composed of Structure From Motion (SfM) and MVS. Various methods are experimented and the quantitative and qualitative results are shown.

## 2 Related works

### 2.1 Key points Detection

Key point detection aims to extract distinctive feature points from images. SIFT, proposed in [1], is invariant to scale and rotation, and uses the Difference of Gaussian to determine key points. SURF [2] uses a Hessian matrix-based measure, and leverages a distribution-based descriptor. ORB[3] is another feature extractor that first detects FAST points, then leverages the intensity centroid in corner orientation.

### 2.2 Sparse Stereo Reconstruction

Sparse reconstruction obtains a set of sparse 3D points by first recovering the transformation between the two images, then obtains the 3D sparse set of points through triangulation. 8-point algorithm is a classical algorithm that recovers the essential or fundamental matrix from the epipolar constraint [4]. An alternative to the 8-point algorithm is the RANSAC algorithm which is a general parameter estimation approach and is designed to cope with outliers in the input data [5].

### 2.3 Dense Stereo Reconstruction

Dense stereo matching determines the corresponding points of each pixel from two or more images(Multiview Stereo). StereoSGM performs a fast approximation by path-wise optimizations from all directions [6]. It leverages a pixel-wise, Mutual Information (MI)-based matching cost for compensating radiometric differences of input images. [7] is an alternative method based on the sum of absolute differences.

## 3 Method

To obtain the 3D point cloud of the target, we first use the SfM (structure from motion) to calculate the relative position of the stereo image pair. The rotation and translation

are then refined by bundle adjustment and are used to rectify the images while obtaining the projection matrix Q. Finally, through the global or semi-global matching, we can yield the depth map which leads to the 3D point cloud via the matrix Q.

## 3.1 Structure from motion

### 3.1.1 Feature point detection and matching

We extracted the key points using SIFT, SURF, and ORB [1, 2, 3], then applied Brute-Force Matcher and Flann matcher to match the points based on their similarities. To handle mismatches, we utilized the ratio-test to filter out the under-qualified matches [1].

### 3.1.2 Fundamental matrix and Essential matrix

We estimate the fundamental matrix and the essential matrix using the 8-point algorithm and RANSAC:
**8 point algorithm** is based on the epipolar constraint:

$$x_1 F x_2^T = 0$$

Where $x_1$, $x_2$ are the projection of a point onto the two images and are homogeneous coordinates, while $F$ is the fundamental matrix. Given the coordinates of more than 8 pairs of matched points, we can estimate the fundamental matrix by singular value decomposition.

we then obtain the essential matrix from the equation:

$$KFK^T = E$$

Where $K$ is the intrinsic matrix. We make sure the constrain of the essential matrix is fulfilled. Note that the 8-point algorithm is sensitive to the outliers
**RANSAC** (Random sample consensus) helps remove the outliers among the matched points. It iteratively samples points to fit the model until the inliers ratio exceeds a threshold [5]. This inliers ratio is used to evaluate different matching methods in Section 4.1.2

### 3.1.3 Recover pose and apply bundle adjustment

From the essential matrix, we recover the rotation matrix $R$ and the translation matrix $T$. We apply Singular Value Decompose on the essential matrix $E = UDV^T$, Then, four possibilities of R and T are:
1. $T_1 = U[:, 3]$ and $R_1 = UWV^T$
2. $T_2 = -U[:, 3]$ and $R_2 = UWV^T$
3. $T_3 = U[:, 3]$ and $R_3 = UW^TV^T$

4. $T_4 = -U[:, 3]$ and $R_4 = UW^TV^T$
where

$$W = \begin{bmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The set of R and T that make most of the points positioned in front of the cameras are chosen as the rotation and translation between the two images.

We then use bundle adjustment to refine the R and T by optimizing a re-projection error. The error is minimized by the Levenberg-Marquardt method using Ceres [8]:

$$Error = \sum_i \left\| x_{i1} - \pi_{left}\left(T^i_{left} \cdot X_i\right)\right\|^2 + \left\| x_{i2} - \pi_{right}\left(T^i_{right} \cdot X_i\right)\right\|^2$$

$x_{i1}$ and $x_{i2}$ are the coordinate of the key points from the image, $X_i$ is homogeneous representation of the corresponding 3D coordinate, $\pi$ and $T_i$ are the projection factors from the intrinsic matrix and extrinsic matrix.

## 3.2 Multview Stereo (Dense Matching)

### 3.2.1 Image rectification

The image rectification process aligns the parallel epipolar lines to simplify the stereo matching process. In addition, we obtain a projection matrix $Q$, that projects the depth map onto the point clouds.

### 3.2.2 Semi-global matching and point cloud generation

We apply Semi-global Matching (SGM) to generate disparity maps [9]. SGM takes advantages of local matching and global matching. The energy function is:

$$L_r(p,d) = c(p,d) + \min \left\{ \begin{array}{c} L_r(p-r,d) \\ L_r(p-r,d \pm 1) + P_1 \\ \min_{i=d_{\min},\ldots,d_{\max}} L_r(p-r,i) + P_2 \end{array} \right\} - \min_{i=d_{\min},\ldots,d_{\max}} L_r(p-r,i)$$

where the first term is the data term, and the second term enforce the smoothness of the matching while the third ensures that $L_r$ does not exceed the upper limitation. The matching cost under all parallaxes of the pixel is aggregated in one-dimensional cost on all paths (such as 8 or 16) around the pixel, and then all the aggregated values of the one-dimensional costs are added to approximate the two-dimensional optimal.

SGBM (Semi Global Block Matching) and StereoBM from OpenCV are used to generate the disparity maps [6, 7]. And the point cloud is generated from:

$$[X, Y, Z, W]^T = Q \cdot [x, y, disparity(x, y), 1]^T$$

# 4 Results

## 4.1 Quantitative results

### 4.1.1 Keypoint detection

We evaluate three detectors: ORB, SIFT, and SURF [1, 2, 3]. And we extract points from all 24 scenes in the Middlebury Dataset [10]. As shown in Table 1, SURF extracts significantly more points than ORB and SIFT.

| keypoint method | Avg num of points extracted in one scene |
|---|---|
| ORB | 500 |
| SIFT | 3850 |
| SURF | **7708** |

Table 1: Comparison of keypoint detectors

### 4.1.2 Keypoint Matching Algorithm

We match the keypoints by Flann method and brute-force methods, which are followed by the ratio test [1]. The outliers ratios are computed from the RANSAC algorithm. From Table 2, SURF has the highest outlier ratios and the longest runtime. In addition, ORB produces the least key points, while its outlier rate is higher than SIFT.

| keypoint Method | matching method | Outliers Ratio | Avg processing time (s) |
|---|---|---|---|
| SIFT | Flann (KDTree) | **13.77** | 0.634 |
| SIFT | Brutal force | **12.94** | 0.608 |
| ORB | Flann (LSH) | 17.64 | 0.111 |
| ORB | Brutal force | 14.39 | **0.104** |
| SURF | Flann (KDTree) | 21.46 | 0.660 |
| SURF | Brutal force | 20.23 | 0.795 |

Table 2: Ratio of the outliers, and average processing time (in second)

### 4.1.3 Disparity Generation

To evaluate the disparity maps, we use the bad2.0 metric from [10], which is the percentage of the bad pixels with disparity error larger than 2 pixels. We record the average processing time in Table 3 as well, in which SGBM produces a lower bad2.0 score, but it is four times slower than StereoBM.

| Stereo method | average bad2.0 score | average processing time (s) |
|---|---|---|
| SGBM | **46.633** | 0.941 |
| StereoBM | 65.213 | **0.216** |

Table 3: Comparison between SGBM and StereoBM.

### 4.2 Qualitative results

In Figure 1 and 2, we compare the generated disparity maps and point clouds with the ground truth. Our disparity maps and point clouds are noisy, exceptionally in background areas such as a white wall. We hypothesize that SGBM and StereoBM cannot match pixels well in areas where the intensity stays constant.

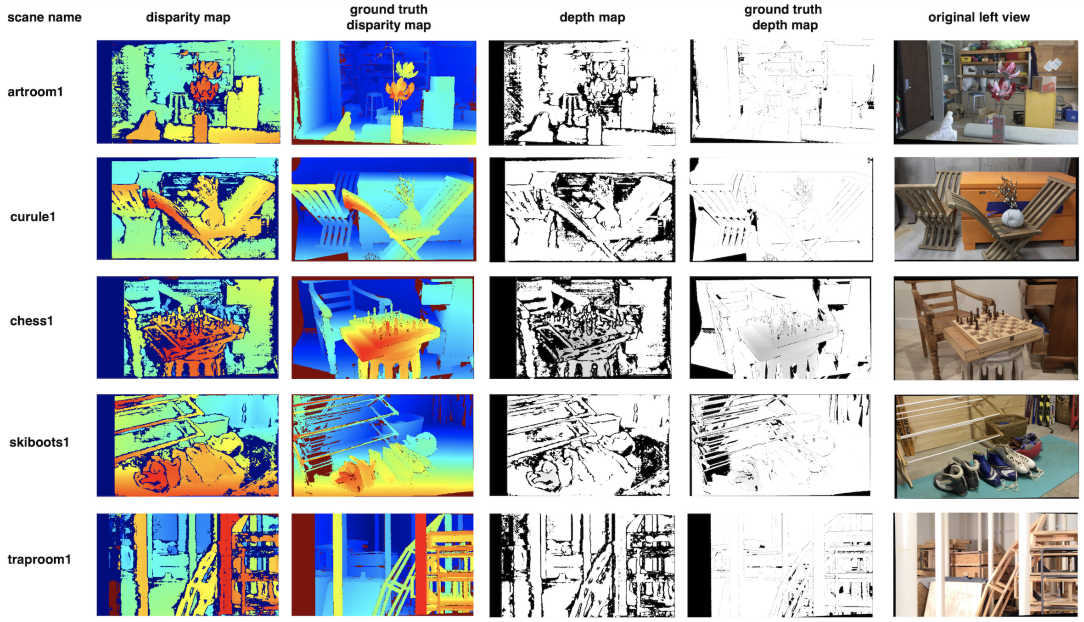


Figure 1: Disparity and depth maps generated compared with the ground truth.

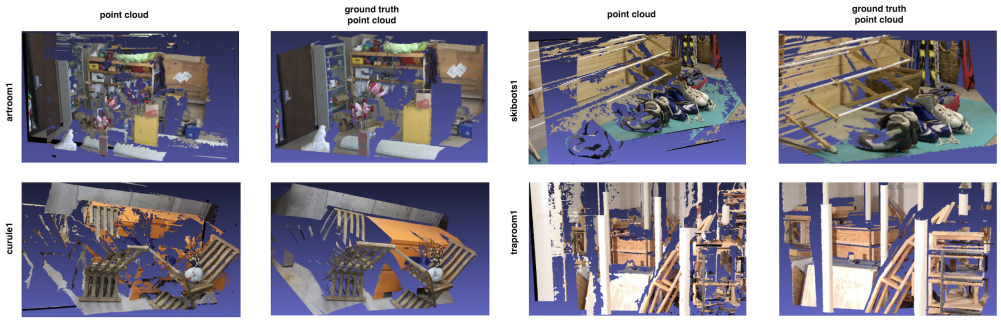


Figure 2: Point clouds generated compared with the ground truth

## 5 Conclusion

We completed a classical stereo reconstruction pipeline that consists of SfM and MVS. We experimented with various methods and compared their performance both qualitatively and quantitatively. The drawback of our method is noisy reconstruction, especially in constant intensity areas. In our future work, we wish to address this issue.

## References


[1] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1, 2, 4

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 1, 2, 4

[3] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, 2011. 1, 2, 4

[4] R.I. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580–593, 1997. 1

[5] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381395, jun 1981. 1, 2

[6] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008. 1, 3

[7] Kurt Konolige. Small vision systems: Hardware and implementation. 1998. 1, 3

[8] Mark K. Transtrum and James P. Sethna. Improvements to the levenberg-marquardt algorithm for nonlinear least-squares minimization, 2012. 3

[9] Heiko Hirschmüller. Semi-global matching-motivation, developments and applications. 2011. 3

[10] Daniel Scharstein, Heiko Hirschmller, York Kitajima, Greg Krathwohl, Nera Nei, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. volume 8753, pages 31–42, 09 2014. 4