Noe Guisset & Ezgi Ozyilkan, `nbg17` & `eo2117`, 01395609 & 01358595

1. Collecting Data

We chose a coffee mug as our textured object to collect data with or without the object in the calibration grid. We recorded 10 images where we changed the camera position between images, and we called this sequence of pictures "**FD**". Secondly, we recorded 12 images by slightly changing the zoom (e.g. a factor of $\sim$ 1.5) and/or rotating the camera whilst keeping its position the same, and we called this sequence of pictures "**HG**". Images are recorded by a Google Pixel 3a mobile phone's 12.2 megapixel rear camera and are of size 4032×3024. The recording took place during daytime within a short time interval to make sure that illumination in the environment is only negligibly variant between images. We experimented with various viewing angles in order to capture the sensory variation in the object of interest (e.g. the beak area of the duck seen in Figure 1). The full image sequences of **FD** and of **HG** can be found in the Section 6.1 and Section 6.2, respectively.

Figure 1: 2 image samples from **HG** dataset.

2. Finding Correspondences

In order to explore *automatic* methods for finding correspondences, we evaluate the performance of three combination of well-known detectors and descriptors: FAST with BRISK, SIFT with SIFT and ORB with ORB. For this end, the *recall* and *precision* values are computed for each image pair containing the object of interest (i.e.the coffee mug) from **FD** and **HG** datasets. In this work, recall and precision between two images are computed as follows:

$$recall = \frac{\#\text{correct matches}}{\#\text{correspondences}}, precision = \frac{\#\text{correct matches}}{\#\text{total positive matches}}. \quad (1)$$

The number of correct matches indicated in Equation (1) is found with Random Sample Concensus (RANSAC) algorithm [5] which eliminates inconsisting matches by selecting the inlier matches and rejecting the outlier ones. The number of correspondences indicated in Equation (1) denotes the number of keypoints in the reference image which are also visible on the query image after the estimated homography matrix of size 3×3 has been applied to them. The number of total positive matches in Equation (1) refers to the number of keypoints in the reference image that have been matched. Note that the recall and precision values indicate the sensitivity of the methods and the relevance of the matched features, respectively[1].

For the floating point based descriptors (e.g. SIFT), we used FLANN matcher [11] with ℓ^2 norm distance. FLANN matcher sorts the best potential matches between similar descriptors based on a distance metric using K-nearest neighbor search. For binary string based descriptors (e.g. BRISK, ORB), we used Brute-Force matcher with Hamming distance. As these techniques provide more matches than we actually need, we finally employed Lowe's ratio test [9] to filter out matches, where we set the threshold to be 0.7. As seen in Table 1, the precision value was significantly higher for **HG** dataset compared to **FD**. We can argue that this is because of the fact that **HG** dataset is obtained by keeping the camera position fixed, therefore making it a relatively easier dataset to obtain correspondences from. FAST+BRISK combination performs best for **FD** dataset (see Figure 11 [in the Appendix]) whereas SIFT+SIFT performs best for **HG** dataset in terms of precision (see Figure 2). ORB+ORB combination performs best in **FD** dataset in terms of sensitivity whereas this combination yields the worst performance in terms of precision in **HG** dataset.

	metric	FAST+BRISK	SIFT+SIFT	ORB+ORB
FD	# matches	17.8	59.9	34.6
	recall	1.30%	1.60%	1.76%
	precision	36.9%	22.6%	25.7%
HG	# matches	43.2	220.8	80.5
	recall	5.14%	17.9%	7.56%
	precision	54.2%	65.8%	42.4%

Table 1: Average number of matches (after filtering out with Lowe's ratio test) and the associated average recall and precision values found using each *automatic* method indicated in the table. Image pairs containing the mug (seen in Figure 1) are chosen from **FD** and **HG** datasets to compute the metrics.

We can also *manually* find correspondences for an image pair by simply clicking on the matching points (see Figure

[1] Sensitivity in this context refers to the true positive rate whereas precision refers to the positive predictive value.

12 [in the Appendix]). However, this manual method is time inefficient and also lacks sufficient precision as we are constrained by the software user interface and by the accuracy of the cursor.

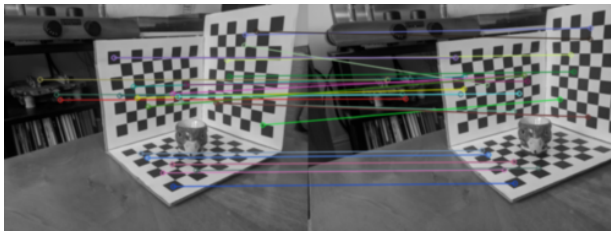

Figure 2: Top 30 matches found with SIFT+SIFT method for an image pair from **HG** dataset. Recall and precision values associated with the matches for the image pair shown in the figure is 35.4% and 79.4%, respectively.

3. Camera Calibration

In order to calibrate the camera used for collecting data in this work, we used the algorithm proposed by Jean-Yves Bouget [2] which is based on the pinhole camera model. The dataset used for the calibration tasks consists of 10 pictures of a 2D calibration target taken from various angles (see Section 6.4.1). Images from **HG** and **FD** dataset are not used as they do not fit the requirements of the built-in calibration functionality of the Computer Vision Toolbox available in MATLAB.

For each image in the calibration dataset (see Figure 13 [in the Appendix]), the corners and keypoints are extracted. A non-linear optimization algorithm is used to minimize the total re-projection error of each keypoint based on the estimated *intrinsic* and *extrinsic* parameters of the camera. The *extrinsic* parameters represent the transformation between the external world and the camera (see Figure 14) whereas the *intrinsic* parameters that represent the internal parameters of the camera itself, such as the focal length, optical center of the camera and lens distortion [10].

Finally, the *extrinsic* (the rotation matrix and the translation vector, i.e. $\mathbf{R}$ and $\mathbf{t}$) and *intrinsic* (i.e. the matrix $\mathbf{K}$) parameters are used to compute the camera matrix, i.e. the matrix $\mathbf{P}$ of size 4×3, as seen in Equation (2).

$$\mathbf{P} = \begin{bmatrix} \mathbf{R} \\ \mathbf{t} \end{bmatrix} \mathbf{K}, \quad \mathbf{K}_{ours} = \begin{bmatrix} 3.26 \times 10^3 & 0 & 0 \\ -12.62 & 3.28 \times 10^3 & 0 \\ 20.25 \times 10^3 & 14.58 \times 10^3 & 1 \end{bmatrix} \tag{2}$$

The pinhole camera model employed in [2] in order to compute the camera matrix, $\mathbf{P}$ in Equation (2), does *not* account for the lens distortion simply because the ideal pinhole cameras do not possess any lenses. Therefore, the *radial* and *tangential* lens distortions are calculated from the keypoint coordinates on the calibration grid. We can then use the computed distortion coefficients to correct the distortion in the images. Note that very little change is observed between the two versions of the image shown in Figure 3. This was expected as modern mobile phones automatically perform post-processing to reduce lens distortion.

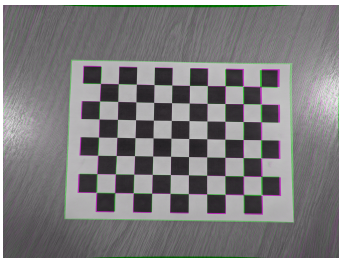

Figure 3: Comparison between the original image (green channel) and its undistorted version (magenta channel).

4. Transformation Estimation

Task1. For this task, we estimate the homography matrix for image pairs retrieved from **HG** dataset. The homography matrix, i.e. $\tilde{\mathbf{H}}$, relates the transformation between two planar scenes such that:

$$\tilde{\mathbf{x}}_{\text{query}} = \tilde{\mathbf{H}} \tilde{\mathbf{x}}_{\text{reference}}, \tag{3}$$

where $\tilde{\mathbf{x}}_{\text{reference}}$ and $\tilde{\mathbf{x}}_{\text{query}}$ correspond to the reference and query homogeneous vectors (or associated image vectors), respectively. Note that although the homography matrix $\tilde{\mathbf{H}}$ is of size 3×3 in our case, it has only 8 degrees of freedom as, by convention, last row's last entry is fixed to be 1, i.e. $\tilde{\mathbf{H}}_{33} = 1$. Note that the matrix $\tilde{\mathbf{H}}$ is also *homogeneous* since only the ratios of its elements are of significance[2]. Due to performing better than other methods in terms of sensitivity and recall metrics shown in Table 1, we used SIFT+SIFT combination to find matches and also applied Lowe's ratio test for filtering purposes. After extracting the locations of the remaining inlier matches obtained through RANSAC in both reference and query images, we compute the homography matrix $\tilde{\mathbf{H}}$ and use $\tilde{\mathbf{H}}$ to project the corners and some keypoints of the reference image onto the query image (see Figure 4 and 15 [in the Appendix]).

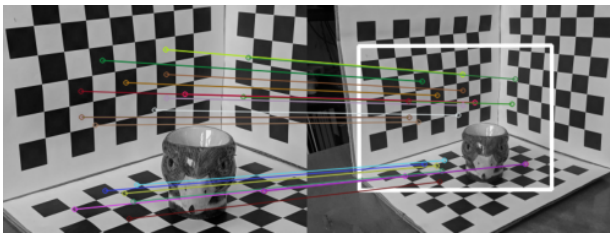

Figure 4: 20 keypoints and their correspondences projected from the reference image (left) using the estimated homography matrix of size 3×3, illustrated for an image pair from **HG** dataset.

Task 2. For this task, we estimate the fundamental matrix for a stereo image pair from **FD** dataset. The fundamental matrix, i.e. $\mathbf{F}$, relates the points in a stereo image pair such that:

[2]Note that this is also the case in homogeneous vectors.

$$\mathbf{x}_1^\intercal \mathbf{F} \mathbf{x}_2 = 0, \quad (4)$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ denotes the corresponding homogenous image coordinates from the stereo pair, and the homogenous matrix $\mathbf{F}$ is of size 3×3 and of rank 2 with 7 degrees of freedom[3]. In Equation (4), $\mathbf{x_1}^\intercal$ lies on the epipolar line $\mathbf{l}_1 = \mathbf{F}\mathbf{x}_2$ corresponding to the point $\mathbf{x}_2$. The line $\mathbf{l}_1$ contains the epipole $\mathbf{e}_1$ for any point $\mathbf{x}_2$ (other than the epipole $\mathbf{e}_2$). Therefore, the epipoles in the stereo image pair are associated with right and left null-spaces of $\mathbf{F}$ such that:

$$\mathbf{F e_2} = 0, \quad \mathbf{F}^\intercal \mathbf{e_1} = 0. \quad (5)$$

We used SIFT+SIFT combination to find matches in the stereo image pair from **FD** dataset. After applying Lowe's ratio test for filtering the matches, we used the least median re-projection error[4] to calculate the matrix $\mathbf{F}$. As seen in Figure 5, it can be inferred that the intersections of the epipolar lines (i.e. the epipoles) in both images lie outside of the visible image frames and therefore, are not shown[5].

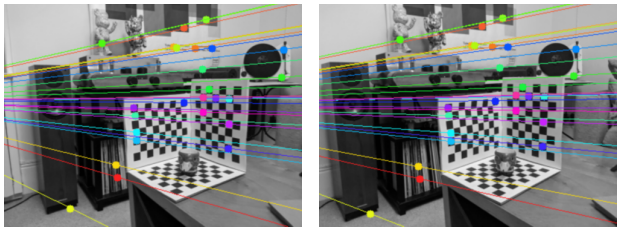

Figure 5: 30 keypoints and their corresponding epipolar lines, illustrated for a stereo image pair from **FD** dataset. In this case, both epipoles lie outside of the visible stereo image pair and therefore, are not shown in the figures.

To estimate vanishing point(s) and the horizon, we used Canny edge detection method [3] (see red curves in Figure 5) followed by the probabilistic Hough Transform [7] to detect straight lines (see blue lines in Figure 6), based on the vanishing point estimation approach proposed in [4].

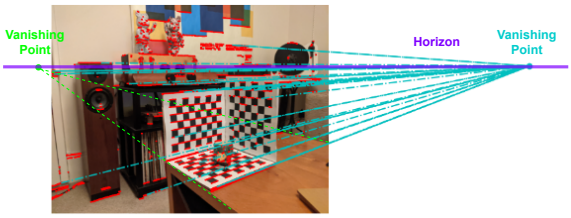

Figure 6: The vanishing points and horizon estimated for the left image of the stereo image pair shown in Figure 5.

5. 3D Geometry

The 3D geometry of the scene can be inferred from a pair of stereo images taken from different perspectives. By comparing the corresponding features, the depth can thus be extracted from the horizontal difference between the rectified images in the stereo pair shown in Figure 7 [6]. The resulting rectified geometry leads to the following relationship between 3D depths Z and disparities d such that:

$$d = f\frac{B}{Z}, \quad x_1 = x_2 + d(x,y), \quad y_1 = y_2, \quad (6)$$

where f and B correspond to the focal length (measured in pixels) and baseline respectively, and x_i, y_i denote the corresponding pixel coordinates on the images in the stereo pair [1]. As seen in Equation (6), the task of extracting depth from a set of images is equivalent to estimating the disparity map $d(x, y)$ once the values of f and B are known.

Task 1. We rectify the stereo image pair shown in Figure 5 such that the epipolar lines become parallel to the horizontal axis and correspondences in both images have identical vertical coordinates [8]. This means that both the images are projected onto a common image plane, hence transforming the disparity calculation from a search on the entire image to a search on a single line. The rectification is performed by finding the epipolar geometry, as explained in Section 4, and applying a projective transformation to both images in order to make the epipolar lines horizontal as discussed (compare epipolar lines shown in Figure 5 and 7 to observe warping effect in the latter figure).

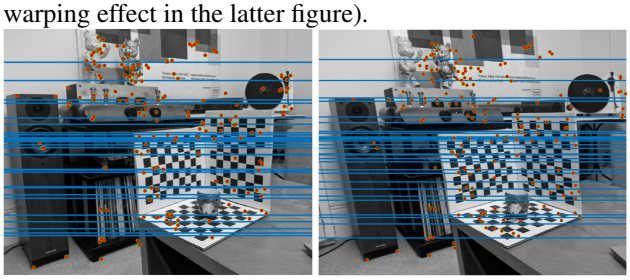

Figure 7: Rectified version of the images in the stereo pair (cf. Figure 5) with epipolar lines being parallel to the horizontal axis. Therefore, epipoles are at infinity in this case.

Task 2. As seen in Equation (6), the amount of horizontal motion, i.e. *disparity*, between the corresponding features in images is inversely proportional to their distance to the camera. To estimate the disparity, we employed Semi-Global Block Matching method from [6], where we set the window size to be 5×5 in order to achieve a reasonable trade-off between good precision and robustness to noise.

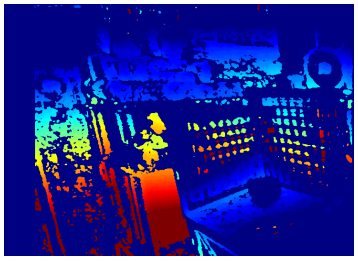

Figure 8: Depth map associated with the rectified version of the images in the stereo pair, shown in Figure 7.

[3] Note that $\mathbf{F}^\intercal$ is the fundamental matrix associated with the stereo image pair of the reverse order, that is $(\mathbf{x_2}, \mathbf{x_1})$ in our case.

[4] The main difference between RANSAC [5] and Least Median of Squares (LMedS) [12] methods is that RANSAC needs a threshold to distinguish inlier matches whereas LMedS only works correctly when the majority of the matches are inlier ones.

[5] See Figure 16 for the epipole provided for the left image in Figure 5.

[6] Horizontal disparity is more popular in the literature than vertical disparity.

6. Appendix

6.1. FD Dataset

See Figure 9.

6.2. HG Dataset

See Figure 10.

6.3. Finding Correspondences

See Figure 11 and 12.

6.4. Camera Calibration

6.4.1 Calibration Dataset

See Figure 13.

6.4.2 Extrinsic Parameters

See Figure 14.

6.5. Transformation Estimation

See Figure 15 and 16.

References

[1] R. Bolles, H. Baker, and D. Marimont. Epipolar-plane image analysis: An approach to determining structure from motion. *International Journal of Computer Vision*, 1:7–55, 2004.

[2] J.-Y. Bouget. *Camera Calibration Toolbox for Matlab*, 2015. (Accessed February 09, 2021). `http://www.vision.caltech.edu/bouguetj/calib_doc/`.

[3] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.

[4] K. Chaudhury, S. DiVerdi, and S. Ioffe. Auto-rectification of user photos. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 3479–3483, 2014.

[5] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, June 1981.

[6] H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 807–814 vol. 2, 2005.

[7] P. V. Hough. Method and means for recognizing complex patterns, Dec. 18 1962. US Patent 3,069,654.

[8] C. Loop and Zhengyou Zhang. Computing rectifying homographies for stereo vision. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 125–131 Vol. 1, 1999.

[9] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[10] MATLAB. Camera calibration.

[11] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *In VISAPP International Conference on Computer Vision Theory and Applications*, pages 331–340, 2009.

[12] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley Sons, Inc., USA, 1987.

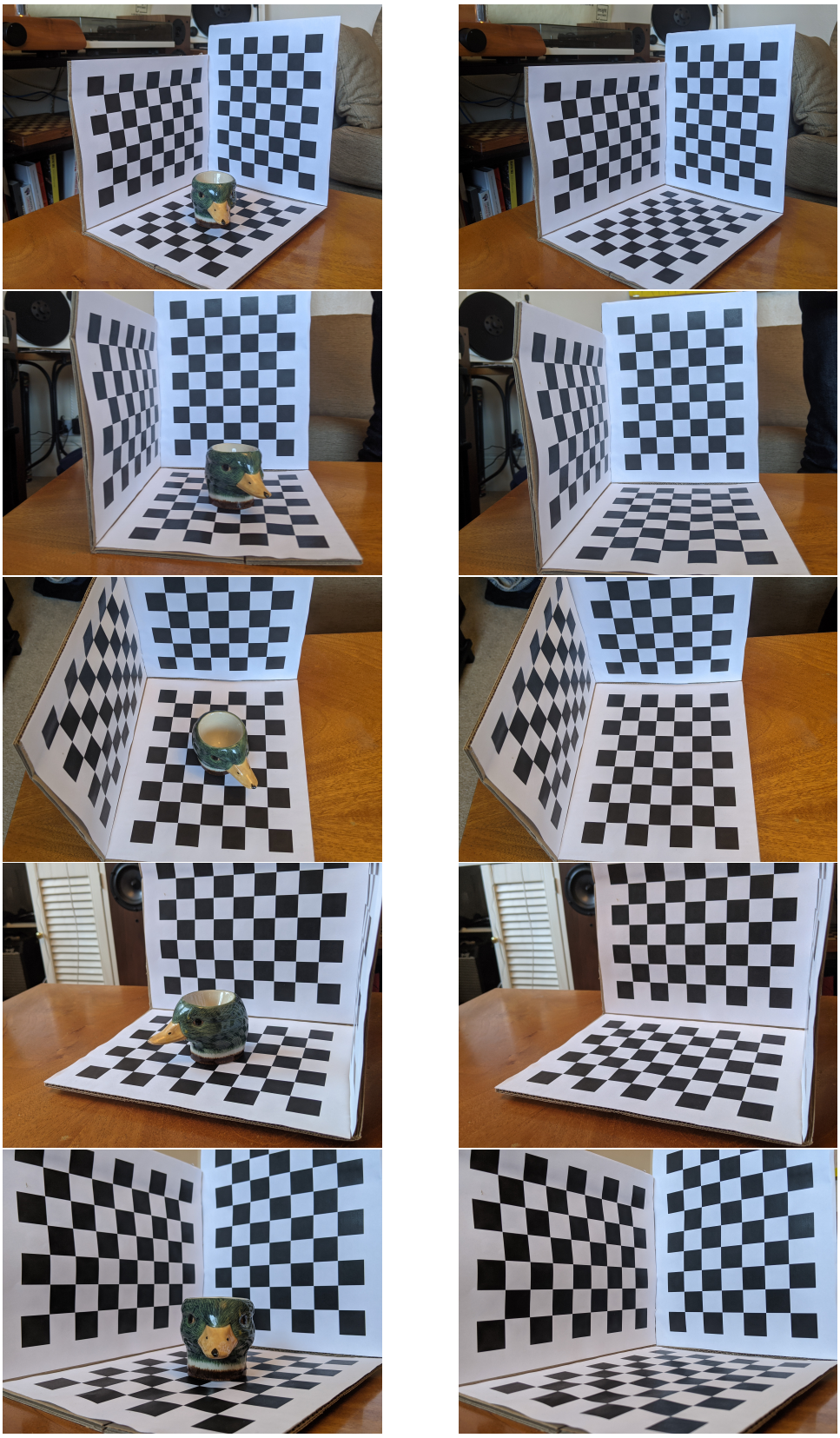

Figure 9: 10 images in **FD** dataset.

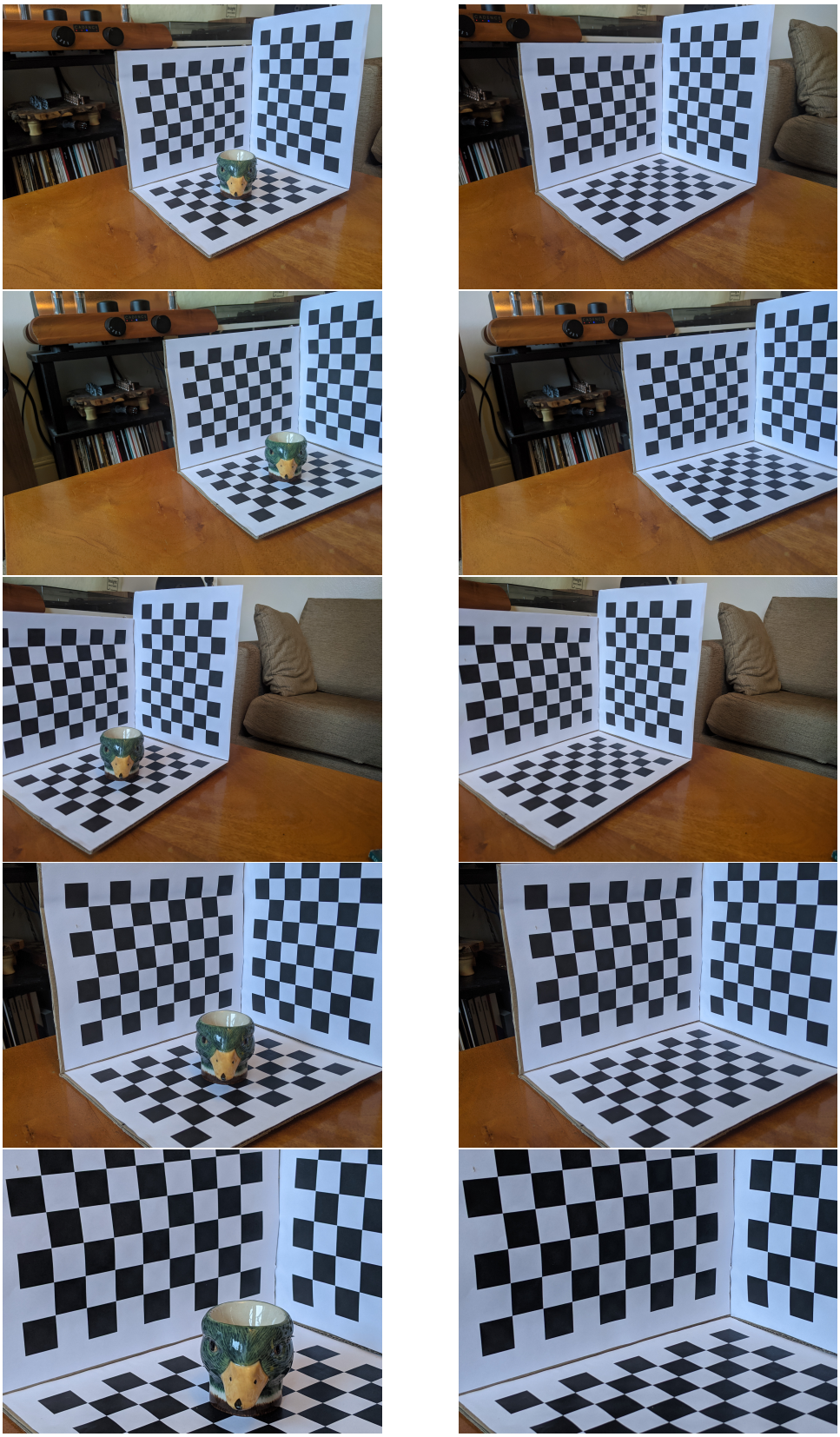

Figure 10: 10 images in **HG** dataset.

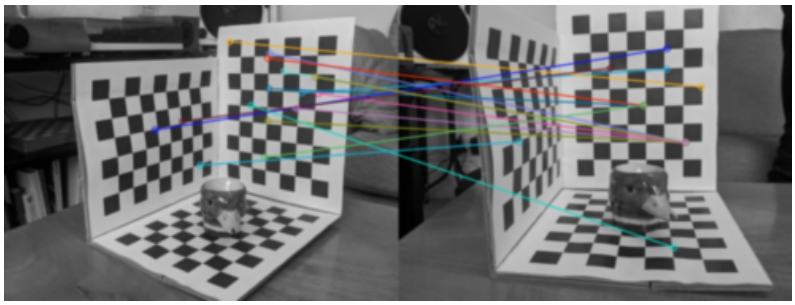

Figure 11: Top 15 matches found with FAST+BRISK method for an image pair from **FD** dataset. Recall and precision values associated with the matches for the image pair shown in the figure is 1.87% and 37.0%, respectively.

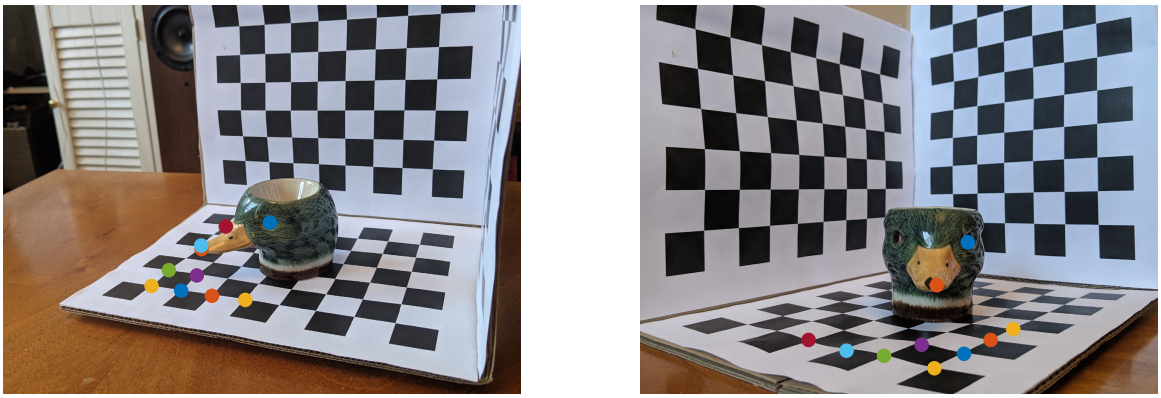

Figure 12: Finding correspondences *manually* for an image pair from **FD** dataset.

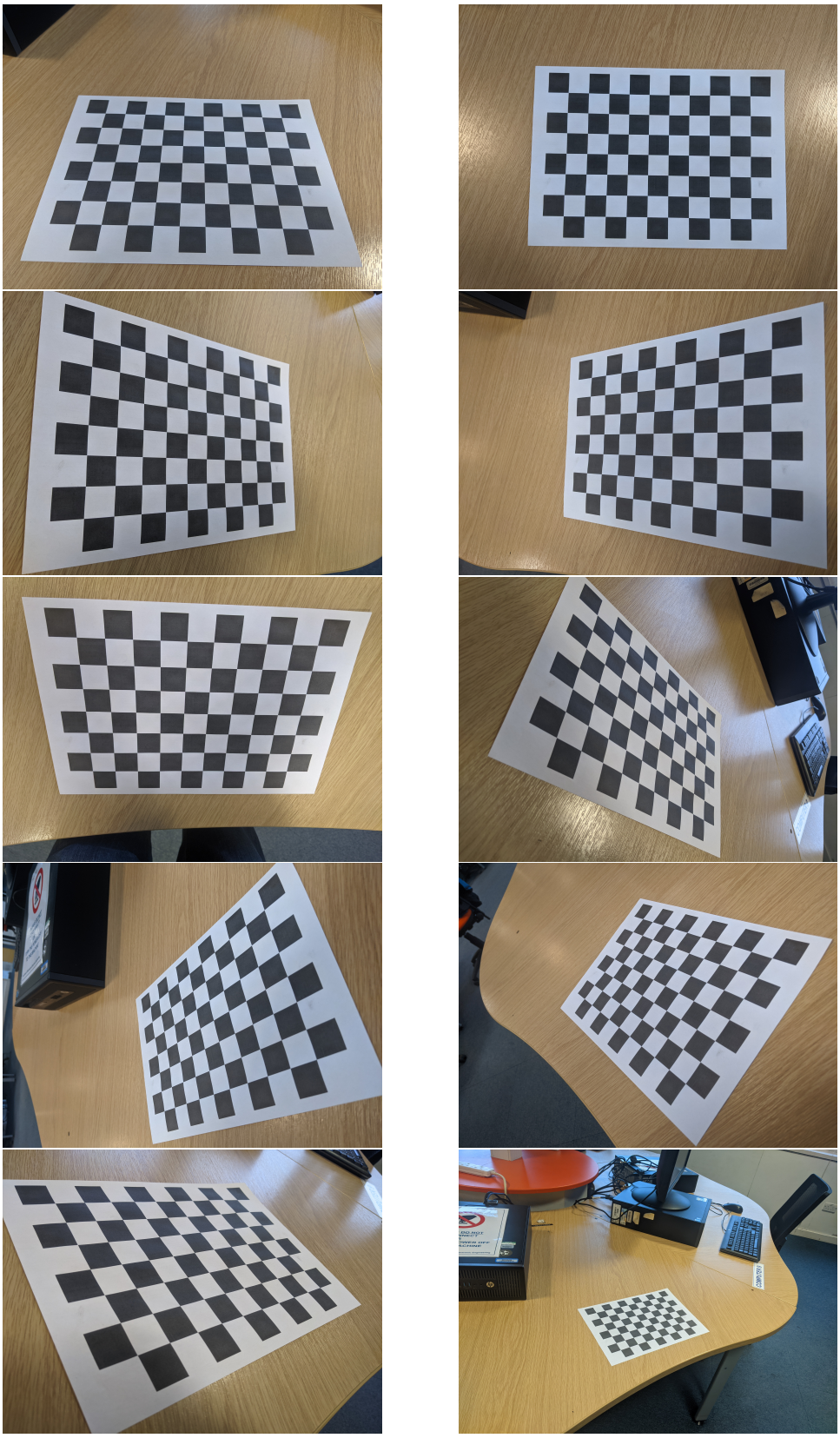

Figure 13: 10 images used for the **camera calibration** task.

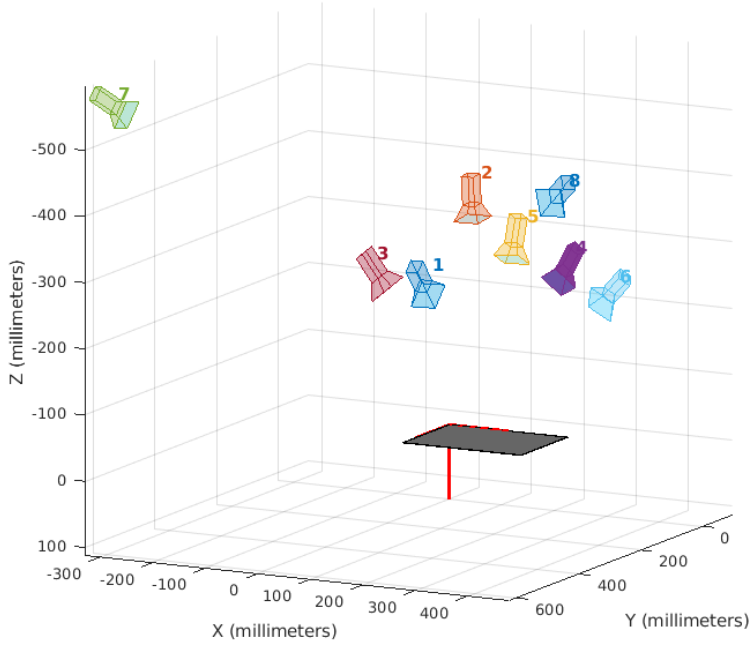

Figure 14: *Extrinsic parameters* extracted from the **calibration dataset** shown on a 3D plot.

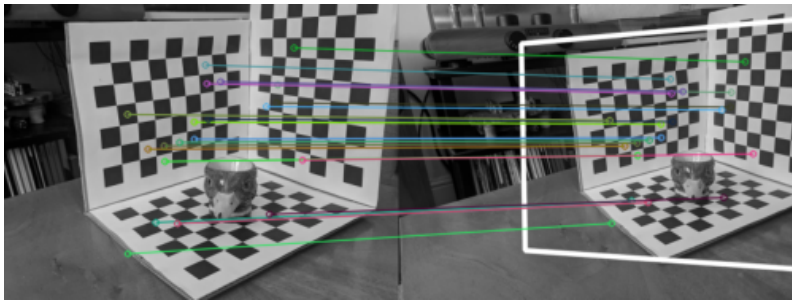

Figure 15: 20 keypoints and their correspondences projected from the reference image (left), illustrated for another image pair from **HG** dataset.

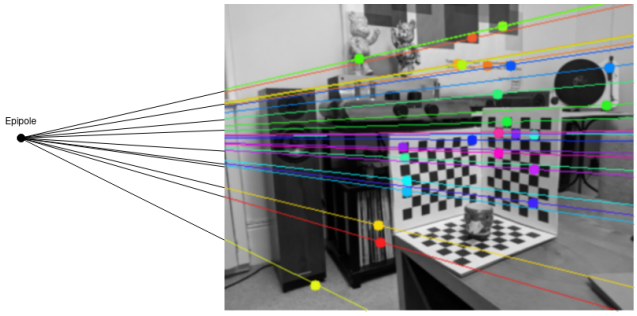

Figure 16: Epipole outside the image frame, shown for the left image of the stereo image pair shown in Figure 5.