Calibrating a Camera and Rebuilding a Scene by Detecting a Fixed Size Common Object in an Image

Levi Franklin

Section 1: Introduction

One of the difficulties of trying to determine information about a 3D scene or the intrinsic properties of a camera based on a 2D image is not knowing the sizes of arbitrary objects in the image relative to the world coordinate frame. Many different techniques are used to try to estimate this information, both automatically and with human assistance. Most calibrations of cameras are done in a laboratory setting and require a fair amount of supplementary supplies and human interaction. [6, 7] This can be cumbersome and inconvenient for people who do not have access to the necessary resources or do not possess the technical skills required. Other more automatic approaches involve using heuristics about the scene and attempting to find distances between points or correlations between multiple images to determine the camera calibration. [5] This approach, however, can be difficult and prone to error. A system that is highly automated without needing difficult to obtain materials would help make camera calibration more accessible.

Another place where lack of knowledge about relative distances in a scene can cause issues is in scene reconstruction. There are many techniques that can be implemented to try to reconstruct a scene from one or multiple 2D images, however, as before, automatic ones are not robust and are prone to error. [6, 7] These can be improved through human assistance, such as picking points to provide a correlation or providing distances between objects, but that can be a difficult task that requires undesired time and effort. A method of gathering information about a scene automatically with very little human interaction can be leveraged to help easily build a more accurate 3D reconstruction.

In this project, I propose that some fixed size objects, when placed in a scene, can be utilized to gain information that would not normally be available. There are many objects that take on very well defined dimensions, shape, and design that are readily available to most people. An example, which we will be focusing on this project, is a US dollar bill. The fact that dollar bills take on a known shape, design, and size can be utilized in some automatic techniques for calibration and scene reconstruction to create more accurate and robust results. Further, because dollar bills are so ubiquitous in society, it is not unreasonable for anyone who is attempting to reconstruct a scene or calibrate an image to come by a dollar for assistance. The goal of this project is to detect dollar bills and use their size and design to assist in recreating aspects of the 3D scene and camera calibration. Though this technique could be applied in a variety of areas, this projects application domain is on mobile devices which are used by individuals who are not skilled in computer vision and where little is known about the camera.

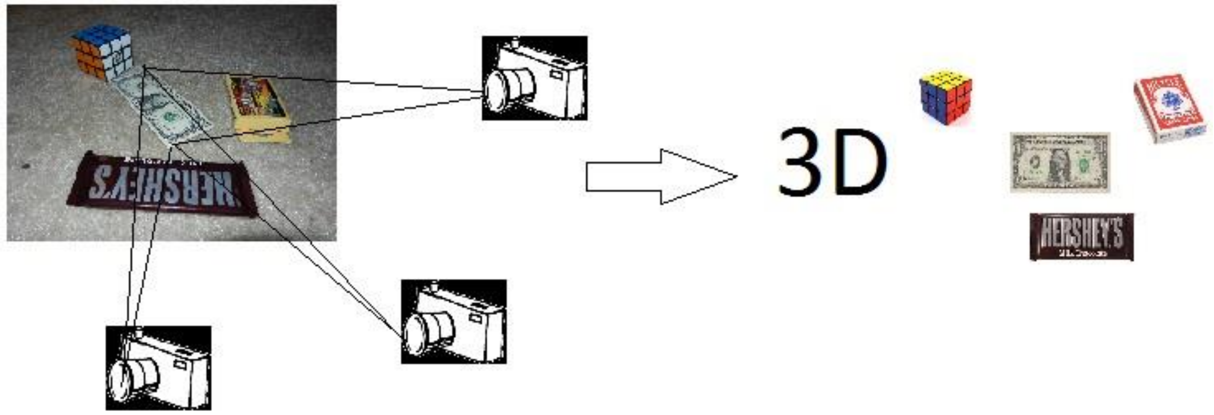

Figure 1: Overview of the technique

Section 2.1: Previous Work

Detecting objects in an image is a common computer vision problem. Many techniques have been developed to assist in object detection. Some techniques involve sliding a "template" of the object across an image. This is quite computationally intensive as well as not robust to changes in illumination, scale, or rotation. More robust techniques such as SIFT and SURF involve finding "keypoints" in an image of your object and matching those to similar keypoints in the image you are searching. [8, 9] Techniques have been developed to find keypoints that are robust to changes in illumination, scale, and rotation. One shortcoming of this technique is that it often does not include information about how the keypoints relate to one another. Spatial matching techniques have also been developed that look for distinct features as in SIFT and SURF, but also keep track of the shape of the object by searching for the spatial relationship between keypoints. [3]

Calibrating a camera is another common task in computer vision that also has been heavily focused on in previous work. A technique involving a "checkerboard" image is the generally accepted method of performing camera calibrations. [5] A checkerboard of known size is placed in a scene and multiple images are taken. The corners of the squares within the checkerboard are known distances from each other. These corners are recovered either automatically or with human assistance and their locations are utilized to estimate the intrinsic parameters of a camera. This technique often requires the printing and careful measuring of a checkerboard image. Other work has been done to automatically detect the corners and estimate the real world coordinates of the checkerboard squares. However, this still usually requires the size of the image be supplied to the calibration algorithm.

The actual reconstruction of a scene is the third problem this project aims to address. This is a difficult process that has had a wide variety of techniques applied to it. Epipolar geometry can be utilized in a scene to attempt to relate points between images. Once accurate and dense point correlations have been found between images, a fundamental matrix can be estimated between them. [7] This gives rise to “epipolar lines” which help further refine corresponding points between two images and can help rectify them. [10] These techniques can be prone to error and often can only solve a scene “up to scale” without a good camera calibration and information about distances within the scene.

Section 2.2: Contributions

My project aims to contribute to all three of the issues mentioned above. Namely, I have a robust method of detecting a dollar bill in a scene, which provides an automated method of calibrating a camera which can assist in 3d scene reconstruction. The system of object detection is able to find dollar bills accurately and robustly in a variety of lighting, at virtually any rotation, and at a variety of scales. I am confident that our algorithm can detect dollar bills with a high level of accuracy taken with typical cameras under most normal operating conditions.

Due to the accuracy of the detection algorithm, a camera calibration is able to be determined from a series of images of the same scene. The dollar matching technique is able to compute a homography which can convert any 2d point in our dollar template to the equivalent point in the search images. This allows a series of points at known distances to be recovered in each image. The prior knowledge of the size of a US dollar allows a correspondence between 2d points in our search images and the true locations in 3d space to be created. This can be utilized, similarly to the checkerboard square locations, to estimate the intrinsic parameters of the camera. Furthermore, a rotation and translation vector relating to the location of the camera in each search image can be estimated which also assists in 3d reconstruction.

Most techniques for reconstructing 3d scenes from a series of 2d images rely heavily on point correspondences between images. These correspondences can be used to rectify two images that were taken from a camera at two different perspectives. However, it can be difficult to know if these correspondences are accurate. The detection of a dollar in a scene immediately provides a large set of correspondences between images. Further, the camera calibration allows 3d points to be triangulated between sets of two images. This calibration helps solve the 3d calibration “to scale” as it has some absolute distance information built in. Further, anything that is on the same plane as the dollar bill can easily be measured which gives a simple, albeit naïve, way of measure distances of objects in an image.

Section 3.1: Technical Summary

This project utilizes some well-known techniques to achieve the contributions it aims to provide. The first problem, the dollar bill detection, utilizes the well-known SIFT (Scale-Invariant

Feature Transform) technique. This provides an abundance of keypoints that can be matched into our search images. However, directly matching sift keypoints leads to many false matches which corrupts our detection. Several filtering/refining techniques are used to create a better set of matches and determine the best homography between our template and the search image.

Once the dollar bill detection technique had been refined enough to be robust, this detection algorithm is applied to a series of images of the same scene. The homographies are stored and used to generate a series of “grids” that correlate to points that are known distances apart from each other. Care is taken to ignore images that don’t have enough valid keypoint matches to constitute a valid dollar detection. Once these are calculated a camera calibration is estimated.

The last step is to move towards a 3d reconstruction. Many different techniques can be combined to help facilitate generating a good reconstruction. It is often helpful to use combinations of two images to generate point clouds for that pair of images. This can then be repeated for all of your images to create a 3d cloud of points that rebuilds the scene. Further, one can use two images to create a disparity map which shows the differences between the two images, and therefore the 3d depth they exhibit. Because the calibration matrices are known, points can also be triangulated between sets of images to further get points in 3d.

Section 3.2: Technical Details

Section 3.2.1: Object Detection

The object detection technique begins by computing SIFT keypoints for the template image. A high resolution image of a dollar is used to compute these keypoints. The SIFT algorithm works by finding features in the image that correspond to corners or other points that are likely to be detectable despite changes in orientation, scale, or illumination. A 128x1 vector is computed which contains information about each keypoint. Due to the high resolution of our image, originally around 15,000 keypoints were discovered in the template.

Next SIFT keypoints are computed on the “search images” that contain an image of a dollar within them. These keypoints are matched to the keypoints calculated in the template image. Each template keypoint is matched to its closest vector in the search image. Additionally, its second nearest match is stored as well. As is mentioned in [9], as long as the ratio of the distance between the first match and the second match is below .8, the match can be considered distinct and therefore a legitimate match. All others are removed as invalid matches.

At this point, the number of valid keypoint matches has significantly been reduced. Many keypoints from the template obviously are not good points as they were determined to be invalid. In order to eliminate bad keypoints, I ran the matching algorithm against a training set of images, and kept track of which keypoints were kept, and which were thrown out for each image. A histogram of the number of images that matched each keypoint was created. I then removed keypoints from my template set that only matched to one or two images. This cut the total number of keypoints being matched by about 4/5. This led to less false matches as well as more rapid matching.

The next goal is to recover a homography from these matches. I found that many times several template keypoints matched to the same point in the search image. This led to the homography often matching all of those keypoints to that one point and thinking it had a good match. In order to combat this, anytime multiple keypoints match to the same point in the search image, the best match is kept and all others are filtered out.

At this point a fairly good set of matching keypoints has usually been recovered. From there a RANSAC method of recovering a homography is utilized. 4 points are chosen at random and are used to generate a homography estimate. Each template keypoint is projected into the search image via the homography and the number of keypoints that are within a certain threshold (15 pixels) of their matching keypoint is recorded. This procedure is repeated until the homography estimate can be confidently considered accurate. This homography is stored for that image and gives the relation for that images dollar detection. If no homography returns enough keypoint matches within the threshold, it is determined the dollar was not found and that image is not used.

Section 3.2.1: Camera Calibration

The benefit of using a fixed size object such as a dollar is evident when doing camera calibration. The size of a US dollar is well established (155.956mm x 66.294mm). This allows one to create a grid as in the checkerboard example. In this project, the dollar is divided up into 40 different points. The width is divided by 8 and the height by 5, creating a grid of 40 points at known distances from each other. This is then projected into each image where the dollar was found by the associated homography. This gives a large set of 2d image to 3d scene coordinate correlations. This is then used to estimate an intrinsic camera calibration matrix as well as rotation and translation matrices for each image in the same manner as [5].

Section 3.2.1: 3d Scene Reconstruction

Multiple techniques can be used to help reconstruct a scene. A simple naïve approach involves using the camera calibration matrix as well as the rotation and translation for each image, to triangulate points from 2d to 3d in sets of two images. Additionally, disparity maps can be created for each of the sets of 2 images that were captured. Each set of images is rectified using the camera calibration matrix, distortion parameters that are estimated from the calibration, and the rotation and translation between the two images.

Once the images are rectified, a disparity map is created triangulating the difference between the two images at each point. This can then be projected to 3d in order to get a reproduction of the scene. When done with all the combinations of images, this leads to a reproduction of the scene in 3d.

Section 4: Experiments

Two image sets were used to test the dollar recognition portion of the algorithm. The first set consists of 23 images of a lone dollar bill. These images were taken at a variety of distances and orientations. The images are taken with normal lighting in the room and did not utilize a flash. Several of the images are somewhat dark or blurry. The second image set consists of 22 images taken of a scene with several other objects also present such as a Rubik's cube and deck of cards. The first 14 images utilize the flash and are very clear and crisp. The last 8 have the same lighting as the first set and are occasionally blurry.

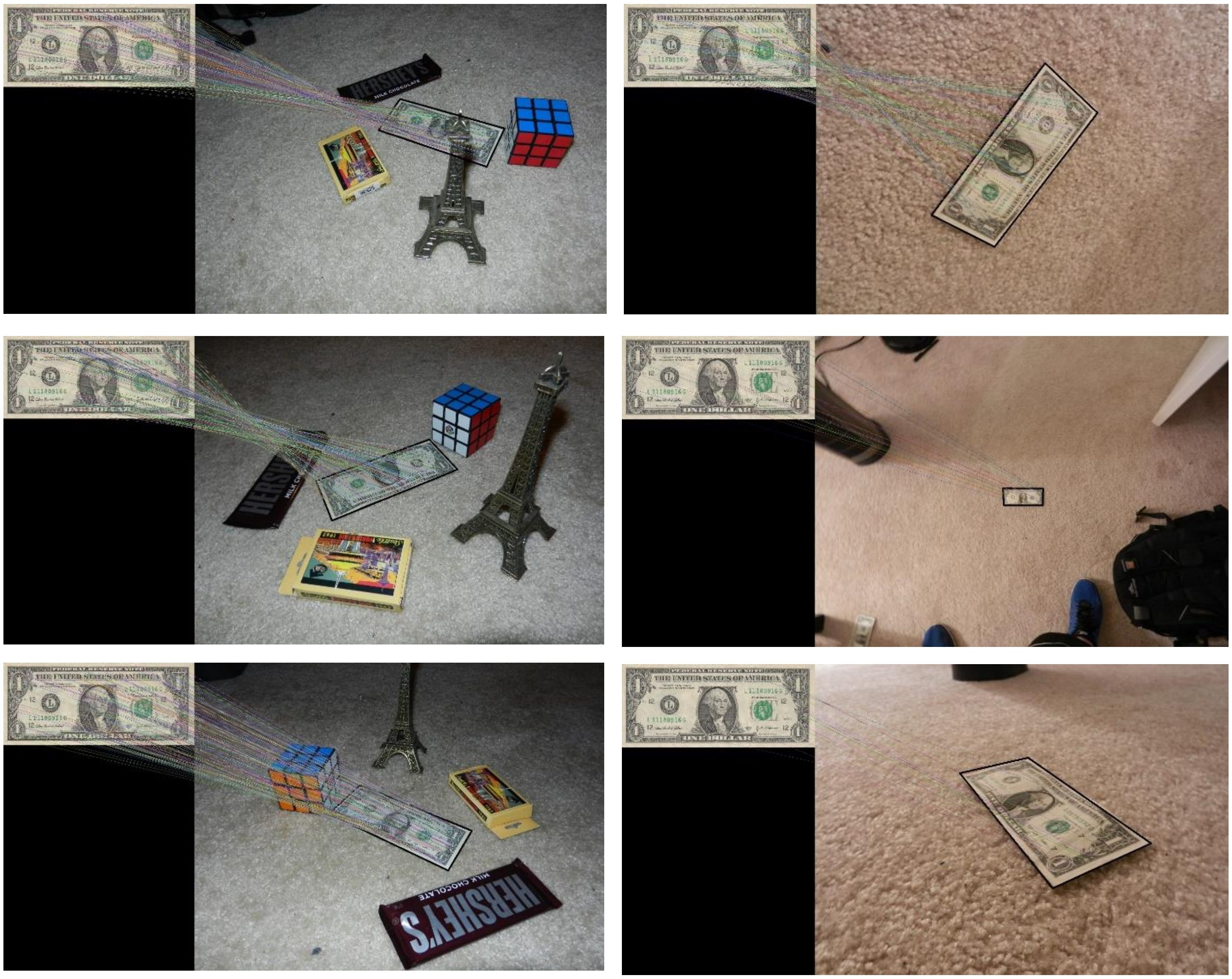

Figure 2: Dataset 2(Left Side) and Dataset 1 (Right side)

To test the dollar finding, I manually found the corners of the dollar in each image to provide a ground truth. Next I computed the corners using our dollar detection algorithm and calculated the average error between the estimated and given corners. I then established a threshold, via visual inspection, of what constitutes a dollar detection and what is a failure.

After observing the results, a threshold of a pixel error over 120 was chosen to constitute a failure. For the first dataset 19/23 images were detected successfully. Among the successful detections, there was an average pixel error of 21.64 pixels. With our image sizes of 4000 x 3000 pixels this is a low error that indicates a very accurate match.

For the second dataset we found that every image in the good lighting portion was found (14/14 images). This section yielded an average error of 12.929 pixels which is again very accurate. When you include the images from that dataset taken under normal room lighting, the dollar was detected in 18/22 images. It is also observed that the images that failed were particularly blurry or taken from unusually low angles.

Once the dollar has been located it is used to generate a set of known 2d to 3d point mappings for each image. This is used in the camera calibration. To test this portion the 14 images that were detected in our second dataset are utilized. The calibration manages to yield a camera calibration matrix and rotation and translation matrix for each image. In order to gauge the accuracy of the calibration, the reprojection error is calculated. A value that is close to zero for the reprojection error reflects a good calibration. After running the calibration 5 times, an average reprojection error of 1.7923 was observed. This is approximately zero and shows that the algorithm achieved a good calibration.

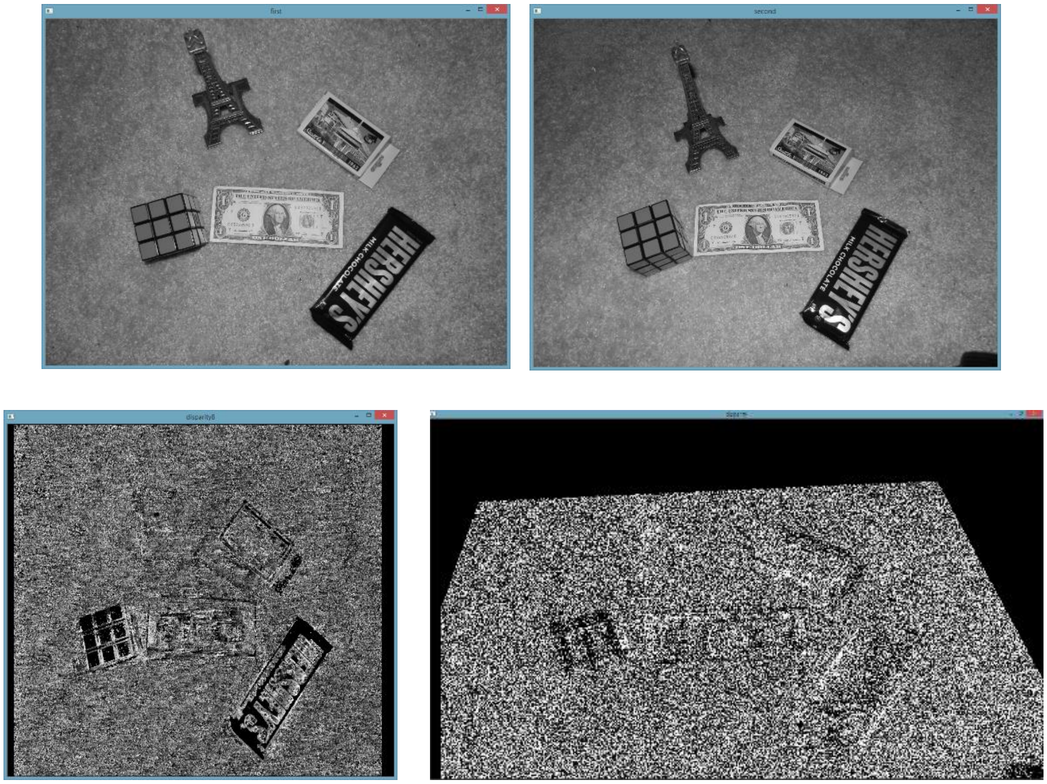

Figure 3: 2 rectified Images (Top) Disparity maps generated from those images (Bottom)

The last experiment conducted involved testing 3d reproduction. In order to test this reproduction a disparity map was created between two images. First the two images are rectified to make their epipolar lines parallel. Then a disparity map between them is generated. Because the images are taken from angles that are quite different from each other it was difficult to achieve a good disparity mapping. Triangulating individual points was also investigated but highly reliable 3d points were difficult to recover. The disparity maps that were recovered can be seen in Figure 3.

Section 5: Conclusions

Overall this project achieved most of its goals. The dollar bill detection algorithm is robust to changes in scale, rotation, or illumination. Further, it is able to detect dollars that are partially occluded with a high level of accuracy. The refinement of keypoints from our template led to faster and more accurate matches. Using RANSAC to recover homographies between the template and search images further refines the matching and allows us to calibrate the cameras.

The camera calibration is generated with an adequate degree of accuracy. Though it is not as robust/accurate as using a checkerboard pattern in a laboratory setting, the benefits of this algorithm are that it is done automatically and doesn't need difficult to acquire assistive tools like checkerboard patterns to do the calibration. The calibration yielded a low reprojection error and the translation and rotation vectors that are recovered assist in 3d reproduction.

Though it is difficult to get a truly 3d representation of the scene, the disparity maps that were created show that 3d depth information can be recovered. If care is taken to adjust the settings and take images from very similar positions, I am confident that this algorithm would allow one to recover a fairly accurate 3d representation of the image. These techniques could easily be adapted to work on a mobile device to allow everyday people to recover information about scenes, such as measuring distances, just by taking an image of a dollar.

Section 6: References

[1] Canny, John, *A Computational Approach to Edge Detection*, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139.

[2] P Viola, M Jones , *Rapid object detection using a boosted cascade of simple features* - Computer Vision and Pattern Recognition, 2001. CVPR …, 2001

[3] S Belongie, J Malik, J Puzicha, *Shape matching and object recognition using shape contexts* - Pattern Analysis and Machine Intelligence, IEEE …, 2002

[4] Pedro F. Felzenszwalb, Daniel P. Huttenlocher, *Efficient Graph-Based Image Segmentation*, International Journal of Computer Vision, 2004

[5] Zhengyou Zhang, *A Flexible New Technique for Camera Calibration,* IEEE Transactions On Pattern Analysis and Machine Intelligence, 2000

[6] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach (2nd Edition).* Prentice Hall, 2011.

[7] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Academic Press, 2002.

[8] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc Van Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346--359, 2008

[9] David G. Lowe. "Distinctive image features from scale-invariant keypoints." IJCV 60 (2), 04

[10] Hartley, R.I., Theory and Practice of Projective Rectification. IJCV 35 2, pp 115-127 (1999)