

Visual Understanding 1 Coursework

Garoe Dorta-Perez
CM50248: Visual Understanding 1
Unit Leader: Peter Hall
Bath University

January 16, 2015

1 Introduction

The main objectives of this unit are to acquire some fundamental knowledge in Computer Vision. In order to gather this skills we will attempt to build a stereo reconstruction system. We refer the readers to Hartley and Zisserman (2004) book on reconstructing a real world scene given several images of it, for an in depth treatise of the topics covered in this report. Using a bottom-up approach the following steps will be taken:

1. **Image convolution:** transforming using kernel matrices.
2. **Feature construction:** detecting scale invariant interest points in images.
3. **Image matching:** using the previous points to match two images.
4. **Stereo reconstruction:** rebuilding 3D points from 2D matched points in stereo images.

2 Convolution

Transforming images has artistic and practical applications, for instance blurring certain parts in a person face to hide defects or detecting borders with differential operators. For simplicity we will restrict ourselves to gray images, defined as two-dimensional $n \times m$ matrices where each element represents the intensity, and is in the range $\{0, 1\}$ where 0 is black and 1 is white. A convolution kernel k , is defined as a $k_1 \times k_2$ matrix, such that $k_1 \leq n$ and

$k_2 \leq m$. When an image I is convoluted with a kernel, a new image I' is generated, such that each element in the second image is calculated as follows:

$$I'(i, j) = I(i, j) * k = \sum_{n=-N}^N \sum_{m=-M}^M k(n + N, m + M) I(n + i, m + j), \quad (1)$$

where $*$ is the convolution operator, i and j are the pixel indices in the image, $N = \frac{1}{2}k_1$, $M = \frac{1}{2}k_2$; since n and m are also indices, the rounding mechanism for N and M becomes an implementation arbitrary criteria. The advantage of this definition lies in the characterization of a flexible framework for several transformations; if most cases it will only be necessary to place the right values in the kernel to achieve the desired transformation. However, we need to specify what happens in the edges of the matrix, for instance copying them in the output image, computing the values circularly, setting them to zero, etc.

An interesting property of convolutions is the *convolution theorem*, which states:

$$g * h = F^{-1}(F(g)F(h)), \quad (2)$$

where g and h are the matrices to be convoluted, F is the Fourier transform and F^{-1} the inverse of the Fourier transform. This property allows to transform the images into the Fourier domain, apply the convolution, transform back and obtain the same results; performing the convolution in the indicated manner is beneficial, as it is more computationally efficient. Below we show some examples of our convolution code, as well as a table comparing the performance of our initial code, our code with the Fourier transform and Matlab's implementation.

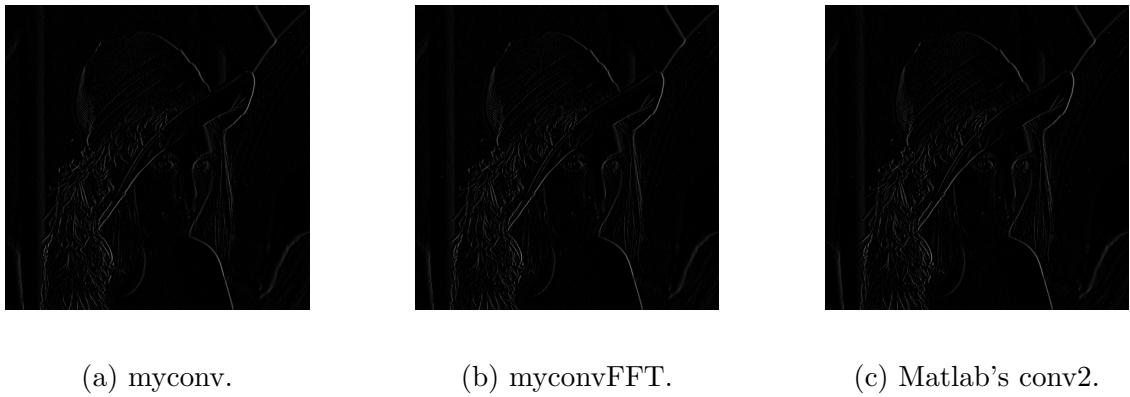


Figure 1: Comparative of the different convolution codes with a horizontal sovel kernel $k = [-1, 1]$.

In Figure 1 we see that the convolution theorem holds, as the output images appear to be identical, and in Table 1 we see that performing the operations in the Fourier domain and transforming back is more efficient.

	myconv	myconvFFT	Matlab's conv2
Time	0.658s	0.04s	0.002s

Table 1: Performance comparative of the convolution methods.

3 Features

Being able to detect image features that are invariant to rotations, translations and changes in illumination or scale is an useful tool, with many applications such as image matching. SIFT features, proposed by Lowe (2004), provide a fairly robust method to detect points in an image with the aforementioned characteristics. A broad overview of the algorithm is shown beneath:

1. **Extrema detection:** maximum and minimum intensity points are detected in a difference-of-Gaussian pyramid built with the input image.
2. **Keypoint refinement:** point centres are interpolated and weak features are rejected.
3. **Orientation assignment:** orientation and strength of each keypoint are calculated.
4. **Descriptor calculation:** a 128 feature vector is computed for every element.

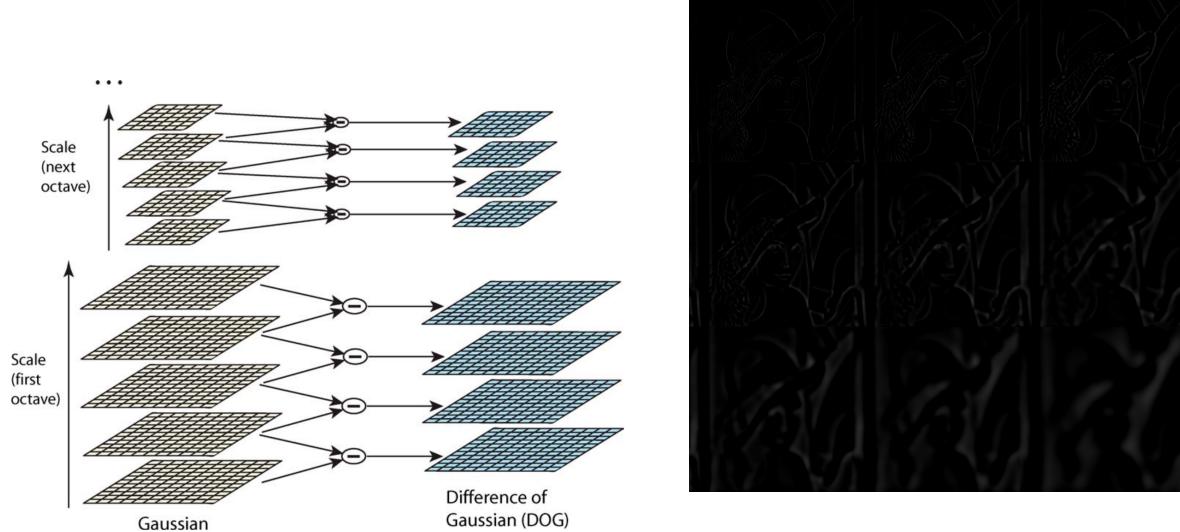


Figure 2: Pyramid of difference-of-Gaussian, Lowe (2004).



Figure 3: Example of a scale of a difference-of-Gaussian.

Extrema detection is computed using a difference-of-Gaussian (DOG) structure, as shown in Figure 2, first a scale of blurred images is constructed. The input image I is convoluted with a Gaussian kernel:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \quad (3)$$

where L is the output image, x and y the pixel indices, and G and Gaussian kernel defined as:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \quad (4)$$

where σ is the standard deviation. Next, the DOG structure is computed as a difference of several L :

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma), \quad (5)$$

where D is the output image and k is a constant multiplicative factor. Note that to provide an efficient implementation, the convolution and difference can be computed in the Fourier domain, using the convolution theorem explained in Section 2. An example on how this DOG structure looks like is shown in Figure 3, where sigma increases from left to right and from top to bottom. The keypoints selected will be those who are a maximum or a minimum in a $3 \times 3 \times 3$ cube region, so in their scale and the ones above and below. Selecting locations using this scheme allows to find keypoints regardless of their scale, as we progress towards the top of the pyramid, the local maximum or minima will be produced by bigger objects in I . In our implementation only one octave of the pyramid is calculated, as shown in Figure 4a.

Keypoint refinement is done with an optimization step which removes weak points and also improves their position. The new keypoint location $\hat{\mathbf{x}}$ will be:

$$\hat{\mathbf{x}} = -\frac{\partial^2 D^{-1}}{\partial \mathbf{x}^2} \frac{\partial D}{\partial \mathbf{x}}, \quad (6)$$

where \mathbf{x} is the previous location and $D(\mathbf{x})$ is approximated with a Taylor expansion in the form of:

$$D(\mathbf{x}) = D + \frac{\partial D^T}{\partial \mathbf{x}} \mathbf{x} + 0.5 \mathbf{x}^T \frac{\partial^2 D}{\partial \mathbf{x}^2} \mathbf{x}. \quad (7)$$

Keypoints are rejected if they are weak or in an edge, any extrema is defined as weak if $D(\mathbf{x}^T) < 0.03$, where $D(\hat{\mathbf{x}})$ is:

$$D(\hat{\mathbf{x}}) = D + 0.5 \frac{\partial D^T}{\partial \mathbf{x}} \hat{\mathbf{x}}. \quad (8)$$

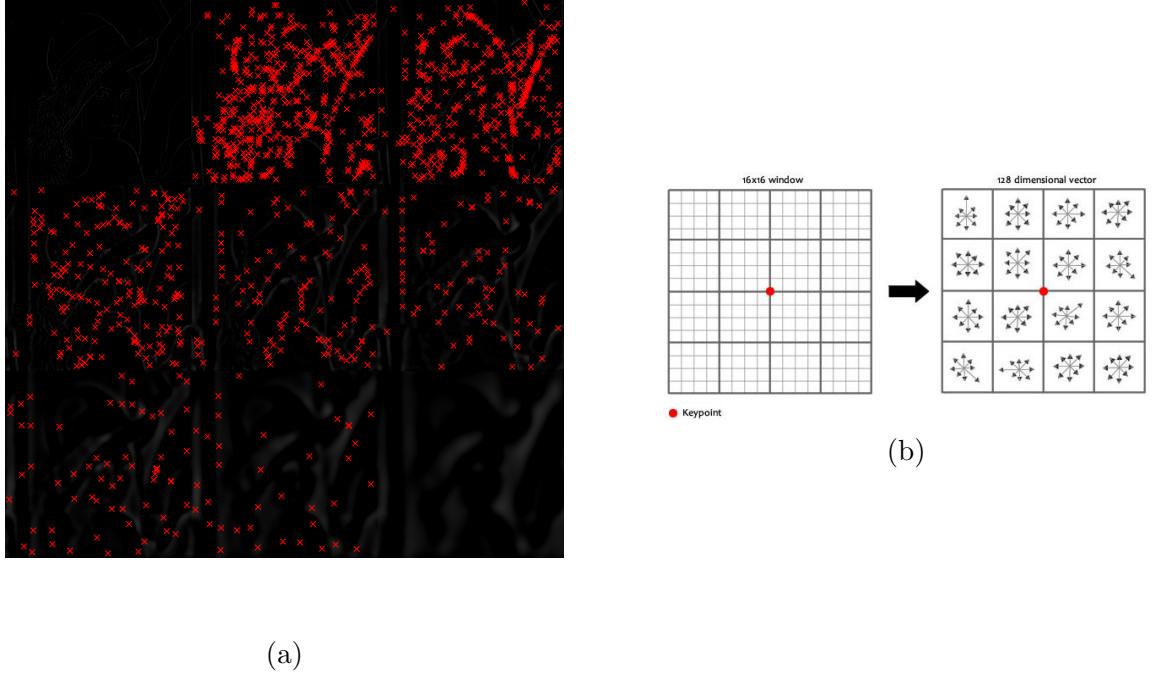


Figure 4: DoG space with simple extrema points detection in (a), and SIFT histogram computation in (b), Gu (2013).

Edge elimination is performed when $R(\mathbf{H}) > f(r)$, where $r = 10$, \mathbf{H} is a 2×2 Hessian matrix and $R(\mathbf{H})$ is defined as:

$$R(\mathbf{H}) = \frac{(D_{xx} + D_{yy})^2}{D_{xx}D_{yy} - (D_{xy})^2}, \text{ where } D_{\alpha\beta} = \frac{\partial^2 D}{\partial \alpha \partial \beta}, \text{ for } \alpha, \beta = x, y, \quad (9)$$

and

$$f(r) = \frac{(r+1)^2}{r}. \quad (10)$$

Orientation assignment measures the point orientation using local image properties. The objective is to obtain the same descriptor for a point regardless of image rotation, i.e. if a keypoint is defined by an object in a picture, we rotate the object and take another picture, the descriptor in the second picture should be identical to the first one. A gradient magnitude $m(x, y)$ and orientation angle $\theta(x, y)$ are computed as follows:

$$\begin{aligned} m(x, y) &= \sqrt{(L(x+1, y, \sigma) - L(x-1, y, \sigma))^2 + (L(x, y+1, \sigma) - L(x, y-1, \sigma))^2}, \\ \theta(x, y) &= \tan^{-1} ((L(x, y+1, \sigma) - L(x, y-1, \sigma)) / (L(x+1, y, \sigma) - L(x-1, y, \sigma))), \end{aligned} \quad (11)$$

where L is chosen to have the same sigma as the keypoint.

Descriptor calculation uses orientation histograms from sample points close to the extrema. For each keypoint a 16×16 region around is selected, and the orientations for each pixel are calculated. 4×4 histograms are computed with 8 bins each; this 8-dimensional values are placed in an array forming a 128 element feature vector for each keypoint, as shown in Figure 4b. Normalizing the vector to unit length improves invariance to illumination changes. In Figure 5, keypoints detected in a sample image by our SIFT implementation and by Matlab’s SURF features, Bay et al. (2006), are shown.

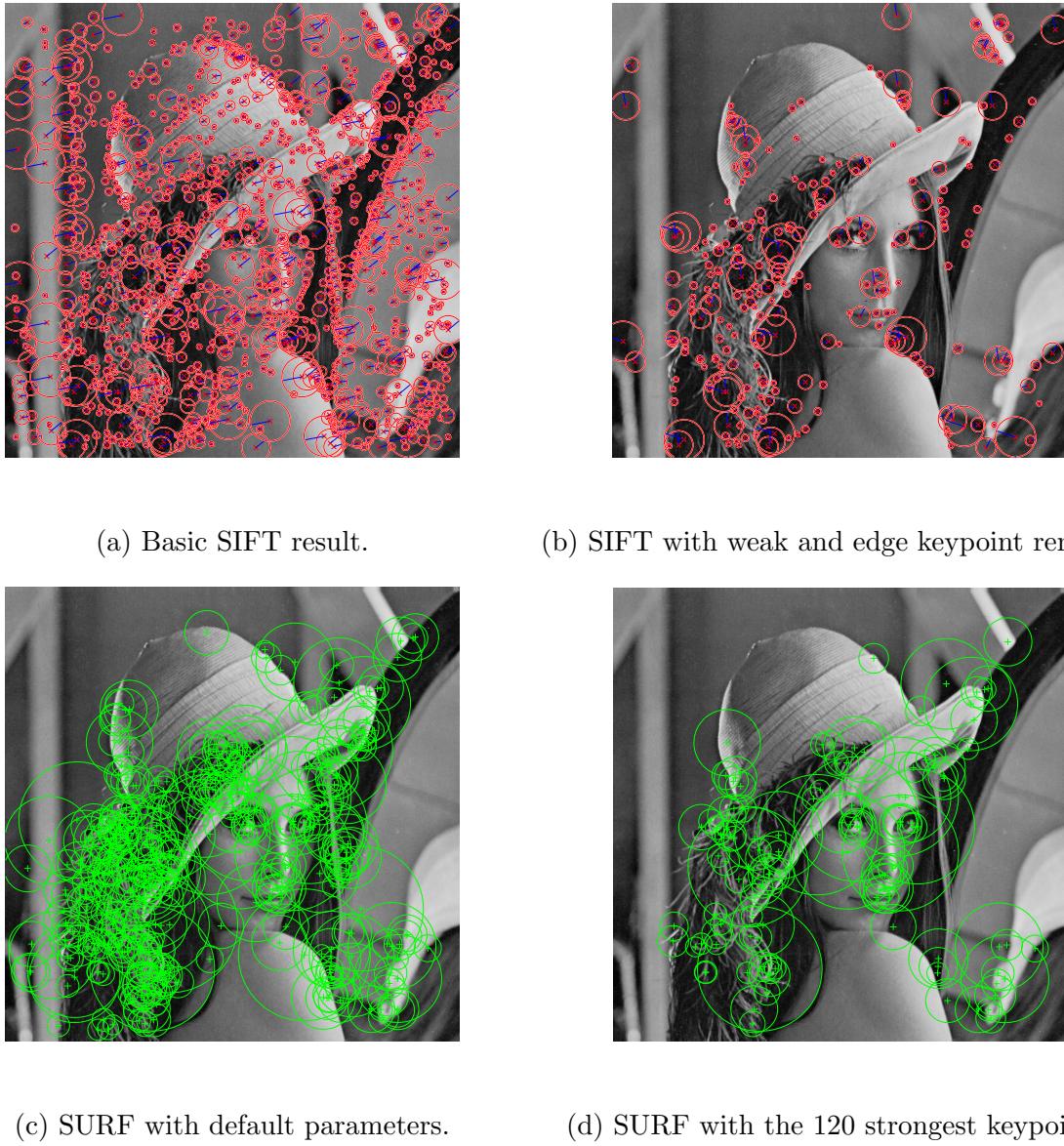


Figure 5: Comparative of our SIFT features with Matlab’s SURF features.

4 Matching

In the image matching area, an image I is to be found under a certain affine transformation in another image I' . Due to camera measurement errors, changes in illumination and other factors, a global optimal solution is computationally expensive. The RANSAC (Random Sample Consensus) algorithm is a simple and fast approach to perform this task, Fischler and Bolles (1981). Despite being originally designed to estimate the parameters that fit a mathematical model to a dataset which contains outliers, it can also be used to estimate the transformation H that matches I and I' . Recently, researchers have proposed more efficient implementations, Dung et al. (2013). A general overview of the algorithm is given below:

Algorithm 1: RANSAC

Data: A pair of images I and I' , number of iterations k , distance threshold t .

Result: Homography matrix H .

$f, f' = \text{detectKeypoints}(I, I');$

$m = \text{matchFeatures}(f, f');$

$H = \text{identity};$

for $i \leftarrow 0$ **to** k **do**

$rp = \text{pickRandomPoints}(m);$

$newH = \text{computeHomography}(rp);$

if $\text{isBetter}(newH, H, t)$ **then**

$| H = newH;$

end

end

To detect and match the keypoints, our implementation uses SURF features. The goal in this stage is to calculate the matrix H that transforms from one image to the other, such that:

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} u' \\ v' \\ w' \end{bmatrix}, \quad (12)$$

where u , v and w are the scaled pixel coordinates in I and u' , v' and w' the scaled pixel coordinates in I' . Collecting H in rows, $H_i = [h_{i1} \ h_{i2} \ h_{i3}]$ for $i = \{1, 2, 3\}$ and $\mathbf{u}' = [u' \ v' \ w']$ we can rewrite Equation 12 into:

$$u = H_1 \mathbf{u}', v = H_2 \mathbf{u}', w = H_3 \mathbf{u}',$$

enforcing $w = 1$,

$$u^* = \frac{H_1 \mathbf{u}'}{H_3 \mathbf{u}'}, v^* = \frac{H_2 \mathbf{u}'}{H_3 \mathbf{u}'}, \text{ hence}$$

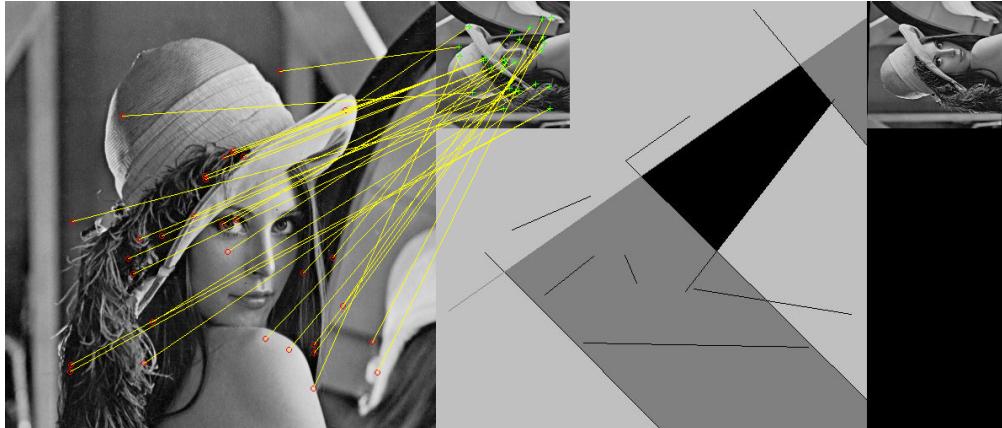
$$u^* H_3 \mathbf{u}' - H_1 \mathbf{u}' = 0, \quad (13)$$

$$v^* H_3 \mathbf{u}' - H_2 \mathbf{u}' = 0, \text{ thus}$$

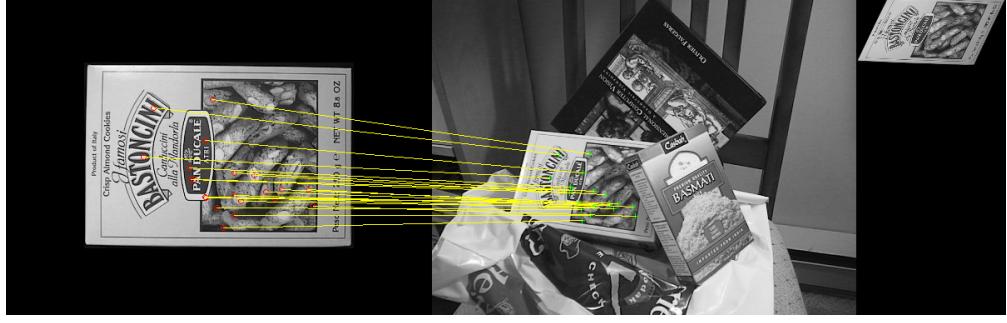
$$\begin{bmatrix} -\mathbf{u}' & \mathbf{0} & u^* \mathbf{u}' \\ \mathbf{0} & -\mathbf{u}' & v^* \mathbf{u}' \end{bmatrix} \begin{bmatrix} H_1^T \\ H_2^T \\ H_3^T \end{bmatrix} = \mathbf{0}^T,$$

where $\mathbf{0} = [0 \ 0 \ 0]$. As shown above, each match provides two equations to solve H . Since we enforce the scale to be homogeneous, $h_{33} = 1$ is no longer an unknown. For eight unknowns, eight equations are needed, so four points in each image are to be matched. In matrix notation, we will stack all the equations in a matrix A , such that $A\mathbf{h} = 0$, where \mathbf{h} is H reshaped into a column vector. Singular value decomposition of A can be computed to solve the system, $A = UZV^T$, where U and V are unitary, Z is rectangular diagonal and $\mathbf{h} = V_{i9}$ for $i = \{1, 2, \dots, 9\}$.

The metric to define if a $newH$ is better than a certain H is double; each matched keypoint in I is transformed with $newH$, the euclidean distance d_i between the transformed keypoint and its match in I' is calculated, and if it is below a threshold t , we count it as a good match. $newH$ is considered an improvement over H if it generates more good matches, or the same number although with lower overall D , such that $D = \sum d_i$. Some results obtained with our code are shown in Figure 6.



(a) A synthetic test.



(b) Matching a box, images taken from Lowe (2004).

Figure 6: Samples of H computation, left is I , centre is I' and right is $I * H$. Keypoints matching between input images is shown for clarity.

5 Reconstruction

3D stereo reconstruction from a pair of cameras involves three steps:

1. **Camera calibration:** compute camera model parameters.
2. **Fundamental matrix:** estimate an F matrix such that $\mathbf{u}'^T F \mathbf{u} = 0$, where \mathbf{u} and \mathbf{u}' are corresponding points in two images.
3. **Reconstruct 3D points:** once constrained by F , triangulate points position from the two images.

We will model our cameras using the pinhole camera model, in this framework a world 3D point $\mathbf{x} = [x \ y \ z \ 1]$ is projected into a 2D image point $\mathbf{u} = [u \ v \ 1]$ using the following equations:

$$\mathbf{u}^T = P\mathbf{x}^T, \quad (14)$$

$$\mathbf{u}^T = K[R|\mathbf{t}]\mathbf{x}^T, \quad (15)$$

where K is

$$\begin{bmatrix} f_x & 0 & p_x \\ 0 & f_y & p_y \\ 0 & 0 & 1 \end{bmatrix} \quad (16)$$

with f_x and f_y are focal length parameters and p_x and p_y are the coordinate of the principal point, P is a 3×4 projection matrix, R is a 3×3 rotation matrix, t is a translation column vector, making $[R|t]$ a 3×4 matrix. Note that \mathbf{x} and \mathbf{u} have a homogeneous coordinate. K is commonly referred as camera intrinsics matrix, $[R|t]$ is called camera extrinsics matrix, and the estimation of their elements is known as **camera calibration**.

Equation 14 can be solved following equivalent steps as to those used in Equations 12 and 13, although with an extra coordinate in \mathbf{x} , and regarding P as equivalent to H with an added column. Therefore we can again stack several points in a matrix A , yet the increased number of unknowns determine that six points are needed to estimate P ,

$$A\mathbf{p} = 0, \quad (17)$$

where \mathbf{p} is P reshaped into a column vector. In the aforementioned technique, the use of a calibration cube is implicit in order to gather the six input points needed. Nevertheless, there are more flexible techniques, such as Zhang (2000) approach to calibrate cameras using at least two images of a calibration plane taken at different camera orientations.

Once we have solved the aforementioned equation with SVD, we need to decompose P into $K[R|t]$. Since K is an upper triangular matrix, we can use an an RQ decomposition on P , where R is upper triangular and Q is orthogonal, Francis (1961). For the decomposition only the first three columns in P will be used, and $\mathbf{t} = K^{-1}P_4$, where P_4 is the last column in P .

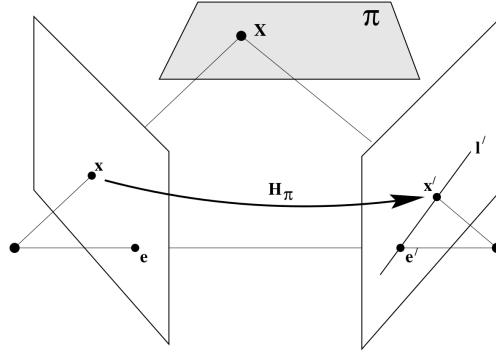


Figure 7: A point \mathbf{x} in the left image is mapped via a plane π to a point \mathbf{x}' in the right image, Hartley and Zisserman (2004).

Epipolar lines and the **fundamental matrix** F are useful tools in stereo reconstruction. In Figure 7, the geometrical relationship of the projection of a 3D world point \mathbf{X} in two camera images is shown. We define \mathbf{e} and \mathbf{e}' as the epipolar points (epipoles), F is a matrix that satisfies $\mathbf{u}^T F \mathbf{u} = 0$, where $\mathbf{u} = [u \ v \ 1]$ and $\mathbf{u}' = [u' \ v' \ 1]$ are matching points in the left and right images respectively, and \mathbf{l}' is a line such that $\mathbf{l}' = F\mathbf{u}'$.

To calculate F the approach in Equations 12 and 13 can be used again. However, in the equation $A\mathbf{f} = 0$, A will have following configuration:

$$\begin{bmatrix} u_1u'_1 & v_1u'_1 & u'_1 & u_1v'_1 & v_1v'_1 & v'_1 & u_1 & v_1 & 1 \\ \vdots & \vdots \\ u_8u'_8 & v_8u'_8 & u'_8 & u_8v'_8 & v_8v'_8 & v'_8 & u_8 & v_8 & 1 \end{bmatrix} \begin{bmatrix} F_1^T \\ F_2^T \\ F_3^T \end{bmatrix} = \mathbf{0}^T, \quad (18)$$

where the \mathbf{f} is F reshaped in a column vector; F can be estimated using SVD.

Since we have calculated all the necessary parameters for each camera and their relationship, we can now proceed with stereo reconstruction. An implementation in pseudocode is presented below:

Algorithm 2: Stereo Reconstruction

Data: A pair of images I and I' , projection matrices for two cameras P and P' .
Result: Cloud of 3D points \mathbf{X} .

```

 $f, f' = \text{detectKeypoints}(I, I');$ 
 $m = \text{matchFeatures}(f, f');$ 
for  $i \leftarrow 1$  to  $\text{numMatches}$  do
     $A = \text{constructA}(m(i));$ 
     $X(i) = \text{SVD}(A);$ 
end
displayCloud(X);

```

For this task we have a set of correspondences:

$$\mathbf{u}_i = P\mathbf{x}_i, \quad \mathbf{u}'_i = P'\mathbf{x}_i, \quad (19)$$

where the unknown is the 3D world object position \mathbf{x}_i . Solving for a single point \mathbf{x} , we cluster the rows of P , such that $P_i = [p_{i1} \ p_{i2} \ p_{i3} \ p_{i4}]$ for $i = \{1, 2, 3\}$,

$$u = P_1\mathbf{x}, \quad v = P_2\mathbf{x}, \quad 1 = P_3\mathbf{x}, \quad (20)$$

multiplying the first and second equation by the third,

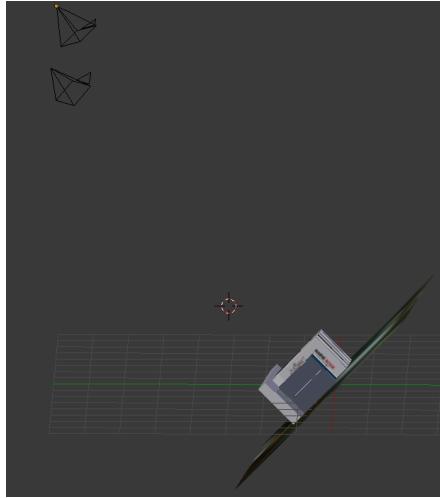
$$\begin{aligned} uP_3\mathbf{x} - P_1\mathbf{x} &\rightarrow uP_3\mathbf{x} - P_1\mathbf{x} = 0, \\ vP_3\mathbf{x} - P_2\mathbf{x} &\rightarrow vP_3\mathbf{x} - P_2\mathbf{x} = 0, \end{aligned} \quad (21)$$

doing the equivalent for the right camera equation, and stacking them in a matrix,

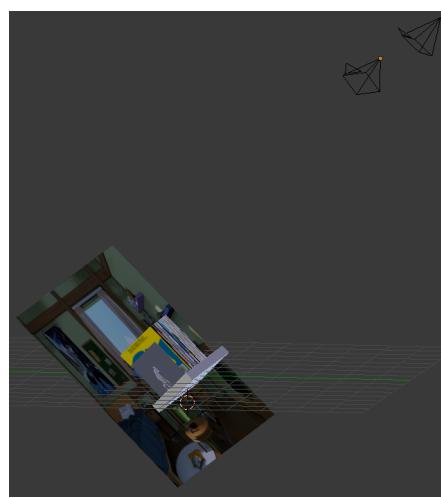
$$\begin{bmatrix} uP_3 - P_1 \\ vP_3 - P_2 \\ u'P'_3 - P'_1 \\ v'P'_3 - P'_2 \end{bmatrix} \mathbf{x}^T = \mathbf{0}^T \rightarrow A\mathbf{x}^T = \mathbf{0}^T. \quad (22)$$

We can calculate \mathbf{x} via singular value decomposition; for each point \mathbf{x}_i a new A_i matrix will be computed and the system will be solved. Our implementation extracts the matches using SURF keypoints, however this produces a sparse reconstruction, as only a small subset

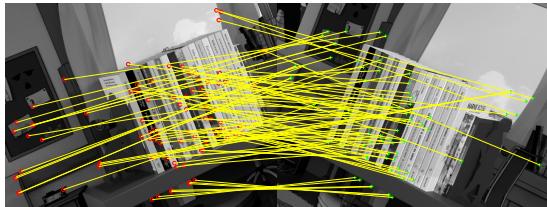
of points in the image is matched. In Figure 8, the work flow of the reconstruction algorithm is shown with an synthetic example. The scene (modelled using Blender) contains some books on a shelf, with a textured plane on the background, based on Archemi (2014). The cameras position can be seen on the upper sections of Figures 8a and 8b. The matching algorithm produces 87 points for this scene, however only 75 are rendered due to some erroneous matches being discarded. Rough forms can be seen in Figure 8e, the two blue circles depict both cameras position; the cloud has three distinctive clusters, the left one shows the points on the background, the centre-top one represents the points on the side-covers of the books, and the lower line-shaped cluster displays the points tracked on the background under the shelf.



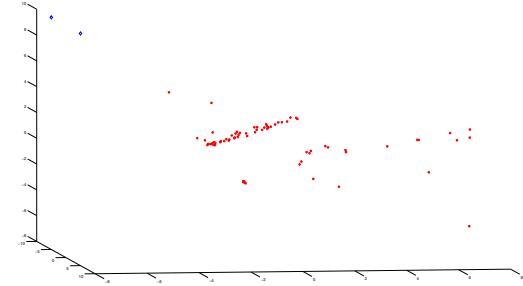
(a) Right view of our 3D scene.



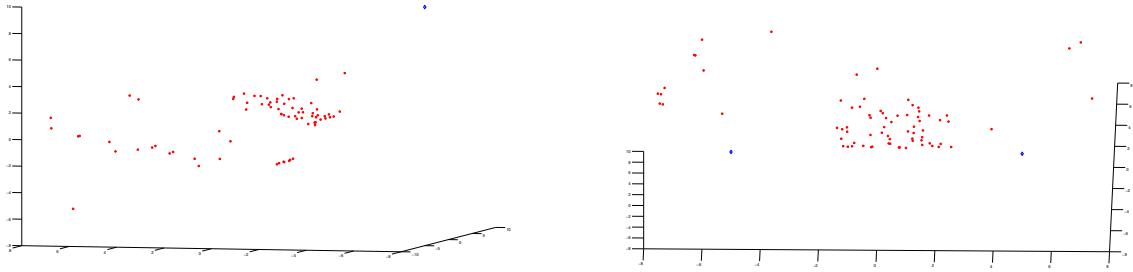
(b) Left view of our 3D scene.



(c) Keypoints matching from each camera view.



(d) Right view of the reconstructed points.



(e) Left view of the reconstructed points. (f) Central view of the reconstructed points.

Figure 8: Reconstructing 3D geometry from a simple scene with books on a shelf.

6 Conclusion and future work

We have presented mathematical derivations and algorithm implementations for convolution, keypoint feature detection, image matching, camera calibration and 3D stereo reconstruction. The next steps would be centred on improving the stereo reconstruction, in order to achieve a denser reconstruction, more points have to be matched, a simple approach to achieve it, is to calculate epipolar lines of matching points, and then performing a neighbour pixel search along the line. Furthermore, to overcome lens distortions in the cameras, and inherent numeric errors in the calibration methods, the epipolar lines can be forced to be horizontal, and recalibration can be done based on matching information.

References

- Archemi (2014). Libri sulla mensola - books. Available at <http://www.blendswap.com/blends/view/75450>.
- Bay, H., Tuytelaars, T., and Van Gool, L. (2006). SURF: Speeded up robust features. In *Computer Vision ECCV 2006*, volume 3951, pages 404–417. Springer Berlin Heidelberg.
- Dung, L., Huang, C., and Wu, Y. (2013). Implementation of RANSAC Algorithm for Feature-Based Image Registration. *Journal of Computer and Communications*, 2013(November):46–50.
- Fischler, M. a. and Bolles, R. C. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6):381–395.
- Francis, J. G. F. (1961). The QR Transformation A Unitary Analogue to the LR Transformation -Part 1. *The Computer Journal*, 4(3):265–271.

Gu, R. (2013). SIFT Histogram picture. Available at <https://courses.cs.washington.edu/courses/cse576/13sp/projects/project1/artifacts/renshugu/RenshuGu1222208.files/image007.jpg>.

Hartley, R. and Zisserman, A. (2004). *Multiple View Geometry in Computer Vision*. Cambridge University Press.

Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Zhang, Z. (2000). A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334.