Computer Vision
CSCI-GA.2272-001
Assignment 3.

November 21, 2019

Introduction

This assignment explores various methods for aligning images and feature extraction. There are four parts to the assignment:

1. Image alignment using RANSAC – Solve for an affine transformation between a pair of images using the RANSAC fitting algorithm. [*30 points*].

2. Estimating Camera Parameters – using a set of 3D world points and their 2D image locations, estimate the projection matrix P of a camera. [*35 points*].

3. Structure from Motion – infer the 3D structure of an object, given a set of images of the object. [*35 points*].

Please also download the `assignment3.zip` file from the course webpage as it contains images and code needed for the assignment.

Requirements

You may perform this assignment in the language of your choice, but Python or Matlab is strongly recommended as they are a high-level languages with much of the required funtionality built-in.

This assignment is due on **Thursday December 19th** at 7pm. Please note that the late policy is as follows: (a) assignments that are late by less than 24hrs will suffer a 10% reduction; (b) those between 24 and 72 hrs late will suffer a 25% reduction and (c) those more 72hrs late will suffer a 50% reduction. You are strongly encouraged to start the assignment early and don't be afraid to ask for help from either the TAs or myself.

You are allowed to collaborate with other students in terms discussing ideas and possible solutions. However you code up the solution yourself, i.e. you must write your own code. Copying your friends code and just changing all the names of the variables is not allowed! You are not allowed to use solutions from similar assignments in courses from other institutions, or those found elsewhere on the web.

Your solutions should be submitted via NYU classes. Please use a single zip file with the filename: `lastname_firstname_a3.zip`. This zip file should contain: (i) a PDF file `lastname_firstname_a3.pdf` with your report, showing output images for each part of the assignment and explanatory text, where appropriate; (ii) the source code used to generate the images (with code comments), along with a master script that runs the code for each part of the assignment in turn.

1 Image Alignment

In this part of the assignment you will write a function that takes two images as input and computes the affine transformation between them. The overall scheme, as outlined in lecture 5 and 6, is as follows:

- Find local image regions in each image
- Characterize the local appearance of the regions
- Get set of putative matches between region descriptors in each image
- Perform RANSAC to discover best transformation between images

The first two stages can be performed using David Lowe's SIFT feature detector and descriptor representation. A Matlab implementation of this can be in found in the `VLFeat` package (`http://www.vlfeat.org/overview/sift.html`). A Python version can be found in the `OpenCV-Python` environment (`http://opencv-python-tutroals.readthedocs.io/en/latest/py_tutorials/py_feature2d/py_sift_intro/py_sift_intro.html`).

The two images you should match are contained in the `assignment1.zip` file: `scene.pgm` and `book.pgm`, henceforth called image 1 and 2 respectively.

You should first run the SIFT detector over both images to produce a set of regions, characterized by a 128d descriptor vector. Display these regions on each picture to ensure that a satsifactory number of them have been extracted. Please include the images in your report.

The next step is to obtain a set of putative matches T. This should be done as follows: for each descriptor in image 1, compute the closest neighbor amongst the descriptors from image 2 using Euclidean distance. Spurious matches can be removed by then computing the ratio of distances between the closest and second-closest neighbor and rejecting any matches that are above a certain threshold. To test the functioning of RANSAC, we want to have some erroneous matches in our set, thus this threshold should be set to a fairly slack value of 0.9. To check that your code is functioning correctly, plot out the two images side-by-side with lines showing the potential matches (include this in your report).

The final stage, running RANSAC, should be performed as follows:

- Repeat N times (where N is $\sim$100):
- Pick P matches at random from the total set of matches T. Since we are solving for an affine transformation which has 6 degrees of freedom, we only need to select P=3 matches.
- Construct a matrix A and vector b using the 3 pairs of points as described in lecture 6.
- Solve for the unknown transformation parameters q. In Matlab you can use the \ command. In Python you can use `linalg.solve`.
- Using the transformation parameters, transform the locations of all T points in image 1. If the transformation is correct, they should lie close to their pairs in image 2.
- Count the number of inliers, inliers being defined as the number of transformed points from image 1 that lie within a radius of 10 pixels of their pair in image 2.
- If this count exceeds the best total so far, save the transformation parameters and the set of inliers.

- End repeat.
- Perform a final refit using the set of inliers belonging to the best transformation you found. This refit should use *all inliers*, not just 3 points chosen at random.
- Finally, transform image 1 using this final set of transformation parameters, q. In Matlab this can be done by first forming a homography matrix `H = [ q(1) q(2) q(5) ; q(3) q(4) q(6) ; 0 0 1 ];` and then using the `imtransform` and `maketform` functions as follows: `transformed_image=imtransform(im1,maketform('affine',H));`. In Python you can use the `cv2.warpAffine` from the `OpenCV-Python` environment. If you display this image you should find that the pose of the book in the scene should correspond to its pose in image 2.

Your report should include: (i) the transformed image 1 and (ii) the values in the matrix H.

2 Estimating the Camera Parameters

Here the goal is the compute the 3x4 camera matrix P describing a pinhole camera given the coordinates of 10 world points and their corresponding image projections. Then you will decompose P into the intrinsic and extrinsic parameters. You should write a simple Matlab or Python script that works through the stages below, printing out the important terms.

Download from the course webpage the two ASCII files, `world.txt` and `image.txt`. The first file contains the (X,Y,Z) values of 10 world points. The second file contains the (x,y) projections of those 10 points.

(a) Find the 3x4 matrix P that projects the world points $\mathbf{X}$ to the 10 image points $\mathbf{x}$. This should be done in the following steps:

- Since P is a homogeneous matrix, the world and image points (which are 3 and 2-D respectively), need to be converted into homogeneous points by concatenating a 1 to each of them (thus becoming 4 and 3-D respectively).
- We now note that $\mathbf{x} \times P\mathbf{X} = 0$, irrespective of the scale ambiguity.

This allows us to setup a series of linear equations of the form:

$$\begin{bmatrix} 0^T & -w_i\mathbf{X}_i^T & y_i\mathbf{X}_i^T \\ w_i\mathbf{X}_i^T & 0^T & -x_i\mathbf{X}_i^T \\ -y_i\mathbf{X}_i^T & x_i\mathbf{X}_i^T & 0^T \end{bmatrix} \begin{pmatrix} P^1 \\ P^2 \\ P^3 \end{pmatrix} = 0 \tag{1}$$

for each correspondence $\mathbf{x}_i \leftrightarrow \mathbf{X}_i$, where $\mathbf{x}_i = (x_i, y_i, w_i)^T$, w_i being the homogeneous coordinate, and P^j is the j^{th} row of P. But since the 3rd row is a linear combination of the first two, we need only consider the first two rows for each correspondence i. Thus, you should form a 20 by 12 matrix A, each of the 10 correspondences contributing two rows. This yields $Ap = 0$, p being the vector containing the entries of matrix P.

- To solve for p, we need to impose an extra constraint to avoid the trivial solution $p = 0$. One simple one is to use $||p||_2 = 1$. This constraint is implicitly imposed when we compute the SVD of A. The value of p that minimizes Ap subject to $||p||_2 = 1$ is given by the eigenvector corresponding to the smallest singular value of A. To find this, compute the SVD of A, picking this eigenvector and reshaping it into a 3 by 4 matrix P.

- Verify your answer by re-projecting the world points $\mathbf{X}$ and checking that they are close to $\mathbf{x}$.

(b) Now we have P, we can compute the world coordinates of the projection center of the camera C. Note that $PC = 0$, thus C lies in the null space of P, which can again be found with an SVD (the Matlab command is `svd`). Compute the SVD of P and pick the vector corresponding to this null-space. Finally, convert it back to inhomogeneous coordinates and to yield the (X,Y,Z) coordinates. Your report should contain the matrix P and the value of C.

In the alternative route, we decompose P into it's constituent matrices. Recall from the lectures that $P = K[R|t]$. However, also, $t = -R\tilde{C}$, $\tilde{C}$ being the inhomogeneous form of C. Since K is upper triangular, use a RQdecomposition to factor KR into the intrinsic parameters K and a rotation matrix R. Then solve for $\tilde{C}$. Check that your answer agrees with the solution from the first method.

3 Structure from Motion

In this section you will code up an affine structure from motion algorithm, as described in the slides of lecture 10. For more details, you can consult page 437 of the Hartley & Zisserman book.

Load the file `sfm_points.mat` (included in `assignment1.zip`). In Python this can be done using `scipy` (`http://docs.scipy.org/doc/scipy/reference/tutorial/io.html`). The file contains a 2 by 600 by 10 matrix, holding the x, y coordinates of 600 world points projected onto the image plane of the camera in 10 different locations. The points correspond, that is `image_points(:,1,:)` is the projection of the same 3D world point in the 10 frames. The points have been drawn randomly to lie on the surface of a transparent 3D cube, which does not move between frames (i.e. the object is static, only the camera moves). Try plotting out several frames and the cube shaped structure should be apparent (the `plot3` command may be useful).

To simplify matters, we will only attempt an affine reconstruction, thus the projection matrix of each camera i will have following form:

$$P^i = \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} M^i & t^i \\ 0 & 1 \end{pmatrix} \tag{2}$$

where M^i is a 2 by 3 matrix and t^i is a 2 by 1 translation vector.

So given $m = 10$ views and $n = 600$ points, having image locations $\mathbf{x}^i_j$, where $j = 1, \ldots, n$, $i = 1, \ldots, m$, we want to determine the affine camera matrices M^i, t^i and 3D points $\mathbf{X}_j$ so that we minimize the reconstruction error:

$$\sum_{ij} ||\mathbf{x}^i_j - (M^i \mathbf{X}_j + t^i)||^2 \tag{3}$$

We do this in the following stages:

- Compute the translations t^i directly by computing the centroid of point in each image i.
- Center the points in each image by subtracting off the centroid, so that the points have zero mean
- Construct the $2m$ by n measurement matrix W from the centered data.
- Perform an SVD decomposition of W into UDV^T.

- The camera locations M^i can be obtained from the first three columns of U multiplied by $D(1:3, 1:3)$, the first three singular values.
- The 3D world point locations are the first three columns of V.
- You can verify your answer by plotting the 3D world points out. using the `plot3` command. The `rotate3d` command will let you rotate the plot. This functionality is replicated in Python within the `matplotlib` package.

You should write a script to implement the steps above. The script should print out the M^i and t^i for the first camera and also the 3D coordinates of the first 10 world points. Cut and paste these into your report.