Fundamental and Essential Matrices (12 p)

a)       What are the Fundamental and Essential Matrices? (3 p)

b)      Describe the algorithm used for computing the Fundamental matrix (3 p)

c)       How do you get the camera motion resolved when the Fundamental matrix is computed? (6 p)

a) Fundamental matrix is a 3x3 squared matrix with 7 degrees of freedom which encodes the epipolar geometry of two views, i.e. given a point in image, multiplying by the fundamental matrix will tell us about the epipolar line in the second view. Essential matrix is like fundamental matrix but requires the points to be aligned to the camera coordinate axis (calibrated camera).

b) The algorithm to compute would be the eight-point algorithm, we have to find the 8 matching features between the two images that we are trying to find the Fundamental matrix, and each match gives a linear equation and thus allowing us to build the constraint matrix A, then we can find f which minimizes ||Af|| using singular value decomposition (SVD) , and getting its least eigenvector of A^TA, finally we need to enforce that F is of rank 2, by finding the matrix F' which minimizes  ||F - F'||. This can be computed by finding the SVD  of F , and let F = UΣV^T, where then F' = UΣ'V^T is the solution.

c) The fundamental matrix F can be written as the essential matrix x′^T F x = 0  ->

x′^T  K′^-T E K^−1 x = 0, where K' and K are the intrinsic matrices of the camera. And E takes the form of  E = [tx]R, so we can compute P = K[I | O] and P' = K' [R|t] and we can replace all of this in the constraint and multiply the translation t by an arbitrary scale factor to get x' = K'RK^-tx + K't/Z.

Image processing (12 p)

a)       Explain how are sampling and aliasing related to CV and what do they mean? (3p)

b)      What does convolution mean in the traditional computer vision (ie when not talking about CNNs)? Explain its factors. (3p)

c)       How is Fourier transform used in CV? (3p)

d)      Where and how are Gaussian pyramids used? How are they built? (3p)

a) Sampling is the process of taking samples of a signal and digitizing it in computer vision, e.g. photons arriving at the active cells on the image sensor are integrated and digitized. Aliasing is the effect of under sampling the signal/image thus resulting in image with jagged edges.

b) Convolution is the mechanism to applying a kernel to the image which results in the a filtered image, it works by masking the image with their neighbourhood of pixels by multiplying the kernel and summing up the products, then outputting the result on the image corresponding to the middle of the kernel.

c) Fourier transform is used in CV to analyze the frequency components of the image, it is also used for analysing what is lost from the sampling of the image, it can also be used to accelerate convolution.  It works by converting the image to its frequency domain in one dimension to analyse/remove high or low frequency of the image/ analysing the orientation of the image from the signal.

d) Gaussian pyramids are used in SIFT for feature detection, the difference of the gaussians images are generated and used to calculate the Laplacian of gaussian approximated to get the local maxima. it is built by convoluting  an image using a gaussian blur, subsequently breaking down the image into increasingly smaller groups of pixels by calculating the weighted average of the neighbouring pixels of the image and scaling it down.

Feature detection and matching (12 p)

a)       What is image gradient? How is it computed? How is it related to feature detection from images?  (2p)

b)      Explain the process of corner detection. What are the pros and cons of using corners as features in feature matching? How are eigenvalues and eigenvectors related to corner detection? (4p)

c)       Explain how is feature matching done in its easiest form. How can the process be improved? (4p)

d)      How is auto-correlation function used in feature detection? (2p)

a) Image gradient is the gradient point in the direction of most rapid increase in intensity, it can be computed by using a sobel filter in the x and y directions, then the resulting image can be used to detect the edge.

b) We take a small window around each pixel in an image, and to identify that the pixel windows are unique, we shift each of them by a small amount in a given direction, and take the sum squared difference of the pixel values before and after the shift, the pixels window where the SSD is large for all the 8 directions are then used to find the local maximum as the corner. The pros is it is invariant to translation rotation and illumination, but it can be slow to compute. The eigenvalues and eigenvectors are used to define shift directions with the smallest and largest change in error.

c) Feature matching is finding the best match of a feature of an image from another image. It can be done by comparing the Euclidean distance of the two descriptors from the two images and finding the descriptor that has the minimum Euclidean distance, in which case it would be a match. It can be improved by also considering the feature distance and compute the ratio distance between the best matching descriptors .

d) It is used in the Harris corner detection to speed up the process as computing exactly the change for each pixel is slow.

Convolutional Neural Networks (12 p)

a)       Where does the term convolution come from into the concept of Convolutional Neural Network and what is the specific feature of a CNN that makes it best suitable for computer vision tasks (2 p)

b)      Explain the main building blocks of a Convolutional Neural Network and their functions (6p)

c)       Select two Computer Vision applications presented during the course where CNNs have been used and explain how they resolve the task. (4p)

a) Convolution in convolutional neural network is similar to the previous concept I have answered, it comes from the applying of the filters to inputs that results in an activation, it can learn of multiple features in parallel on its own for a given image with little preprocessing which makes it best suitable for computer vision tasks.

b) Convolutional layers- Applying the convolution operation to the input layer.

ReLU layers-  Introduce non-linearity to the network for it to train itself via backpropagation.

Pooling layers- Reduce the activation map and speed up the network training by downsampling it.

Fully connected layers- Output consists of a list of probabilities for different possible classes of the image, the label with the highest probability is the class.

c) CNN was used in image classification,  CNN applies many filters to the image, it detects and extracts the feature map of the image and trains itself through forward propagation and backpropagation  which gives the probability of the class as output.

CNN was used in Generative Adversarial Network (GAN) to create synthetic outputs with the style of the input, the generator tries to create random synthetic outputs and discriminator would try to tell the generated output apart from the real outputs. The generator would use discriminator's inference result to learn how to fool it and these processes alternate until they achieve an ideal result.