MPHY0047 – Surgical Data Science

Coursework 2

Release Date:	17 February 2025
Submission Deadline:	10 March 2025
Estimated Coursework Return:	Three weeks after submission
Topics Covered:	Topics 4 – 5
Expected Time on Task:	8 hours

Guidance for Submissions: *Failure to follow this guidance might result in a penalty of up to 10% on your marks.*

I. Submit a single Word//PDF document with questions in ascending order. Explain in detail your reasoning for every mathematical step taken.

II. Insert relevant output (e.g calculations, graphs or figures), and describe them in your document. All calculations, figures and tables must be labelled, showing relevant parameters and units.

III. You will need PYTHON coding to solve the questions. Include **all code** in an Appendix at the end of your document AND as separate PYTHON source files to your submission. Remember to comment on your code, explaining your steps. The submitted source code will be take into account during assessment.

This coursework counts towards 25% of your final MPHY0047 grade and comprises five questions. Question 1 will make up 20% of your CW2 grade, Question 2 will make up 20% of your CW2 grade, Question 3 will make up 30% of your CW2 grade, and Question 4 will make up 30% of your CW2 grade.

On Academic Integrity (Read more about it here)

Academic integrity means being transparent about our work.

- **Research:** You are encouraged to research books and the internet. You can also include and paraphrase any solution steps accessible in the literature and online content if you reference them.
- **Acknowledge others:** We are happy when you acknowledge someone else's work. You are encouraged to point out if you found inspiration or part of your answers in a book, article or teaching resource. Read more about how to reference someone else's work here and how to avoid plagiarism here.
- **Ask *good* questions to your peers:** These include but are not limited to questions like "What is the best mathematical method for this question?", "Should I review any books/materials/videos?", "Which PYTHON function did you use for this problem?", "How was the structure of your PYTHON code?".
- **Be *helpful* and *ethical* when answering questions from your peers:** These include "I think that it would be helpful to review X video, Y page of the slides/notes". "I found this good video online", "I used X PYTHON function, structured that way".
- **Do not share and do not copy:** We expect students **not to share and not to copy** assessment solutions or PYTHON code from their peers, even if partially.
- **Do not publish MPHY0047 assessment material:** We expect students **not to share** MPHY0047 assessment materials at external online forums, including tutoring or "homework" help websites.

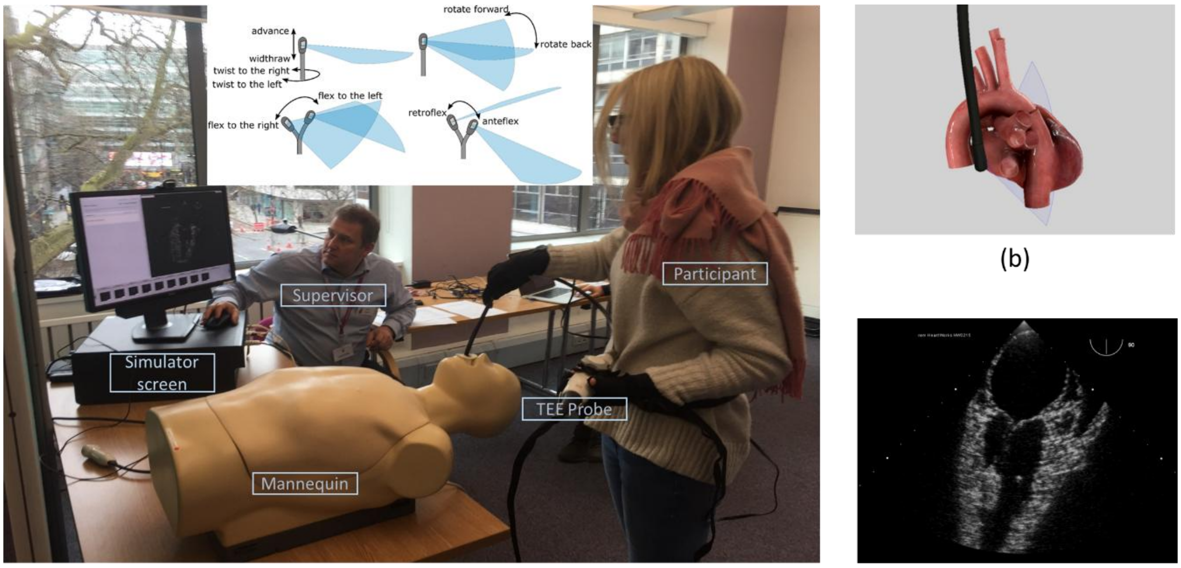
Students found in misconduct can receive a 0 mark in that assessment component and have a record of misconduct in their UCL student register. In some extreme cases, academic misconduct will result in the *termination* of your student status at UCL.

Background

Transoesophageal echocardiography (TOE) is a valuable diagnostic procedure carried out by imaging the heart with an ultrasound (US) transducer, attached to a flexible endoscope (probe). The probe is navigated through the oesophageal lumen adjacently to the heart allowing for high-quality US imaging since with clear depiction of the heart and its operation without interference from skin, muscle or bone tissue, which is the case in the transthoracic surface US. TOE provides a comprehensive depiction of the heart's chambers, valves and blood flow, facilitating hemodynamic assessment. Technological advancements in US scanning transducers, capable of 180° rotation and flexible probe design, enabled the widespread application of TOE, which nowadays is routinely used both perioperatively as a diagnostic and evaluation tool as well as intraoperatively for anaesthesia and hemodynamic management.

Your main task is to investigate image processing features and their potential for estimating manual performance scores of TOE interventionists with different experience.

You will work with a dataset created out of experiments carried out with the HeartWorks TEE simulator (Inventive Medical, Ltd, London, UK). The set-up, shown in Fig. 1a, includes an upper-torso mannequin and a probe with similar capabilities (±180° twist and ±90° flexion) to standard TEE probes. Dedicated VR software renders a high-fidelity 3d model (Fig. 1b) of a beating human heart. An ultrasound detector captures the position and orientation of the US scanning field, used to generate the 2d US image from the heart model (Fig. 1c).



(a) (b) (c)

Fig. 1: (a) A volunteer operating the HeartWorks simulator (inset the HeartWorks TEE probe movements); (b) 3d rendering of the heart model, the probe and US scanning field; (c) the simulated US image.

Good luck !

Problem : Image similarity analysis for determining image quality in Transoesophageal Echocardiography

This study consisted of a single TOE exam in which every participant was asked to capture 10 US views in a specific sequence using the HeartWorks simulator. The 10 cross-sectional views is listed in Fig. 2 (page 5) alongside an illustration of the nominal probe position and orientation with the respect to the 3D heart model, that results to the desired US image [1].

A total of 20 volunteers were recruited and participated in experiments divided into two experience groups according to written information provided with consent. The "experts" group (n=7) comprised solely of anaesthetists having received accreditation and performed more than 500 TOE exams. The "novices" group (n=13) consisted of trainees in the early stages of residency. A consultant anaesthetist performed the same test providing the gold standard (ground truth) images that will be used for comparative analysis.

Manual scoring of the original images was blindly performed by three expert anaesthetists. Each image was assigned two quality scores, one using a standard checklist containing a set of criteria (different for each view) as defined by the ASE/SCA guidelines [2]. Each checklist item was assigned a binary value (0-not met, 1-met) and at the end the percentage (%) of met criteria was calculated. Two examples of checklists for different views are listed in the following table.

Table 1: Checklists used for the ME AV SAX (View 3) and TG 2C (View 7) TEE views.

ME AV SAX (3)
1) 30°-45°rotation
2) AV centred in screen
3) 3 cusps visible
4) imaging plane at level of leaflet tips
5) probe tip appropriately behind LA

TG 2C (7)
1) 85°-95°rotation
2) LA and LV both visible
3) MV visible on right side of screen
4) Post and Ant MV leaflets seen

The second score was assigned based on the mean quality impression of the image and scored on a 0-4 scale by the three evaluators independently. In both scores, the mean value of the three assessors resulted in the final general impression score for the image. Fig. 3 illustrates two examples in the opposite ends of the quality spectrum from two views. The elements in the image that satisfy the criteria in the checklist of each view are designated in yellow circles. The average quality scores are also provided inset. The images on the left are of poor quality and only meet a small number of the checklists' items. For example, the top left ME AV SAX image has the correct probe rotation and visualises the three cusps of the aortic valve. It fails to meet the rest of the criteria. The bottom left image of the TG 2 chamber view only achieved correct probe angulation, but because of inadequate positioning fails to satisfy the rest of the criteria. Consequently, the general impression scores for the left images are also low. On the other hand, the images on the right side are examples of ideally imaged views fully satisfying the respective checklists and achieving full marks in both metrics.

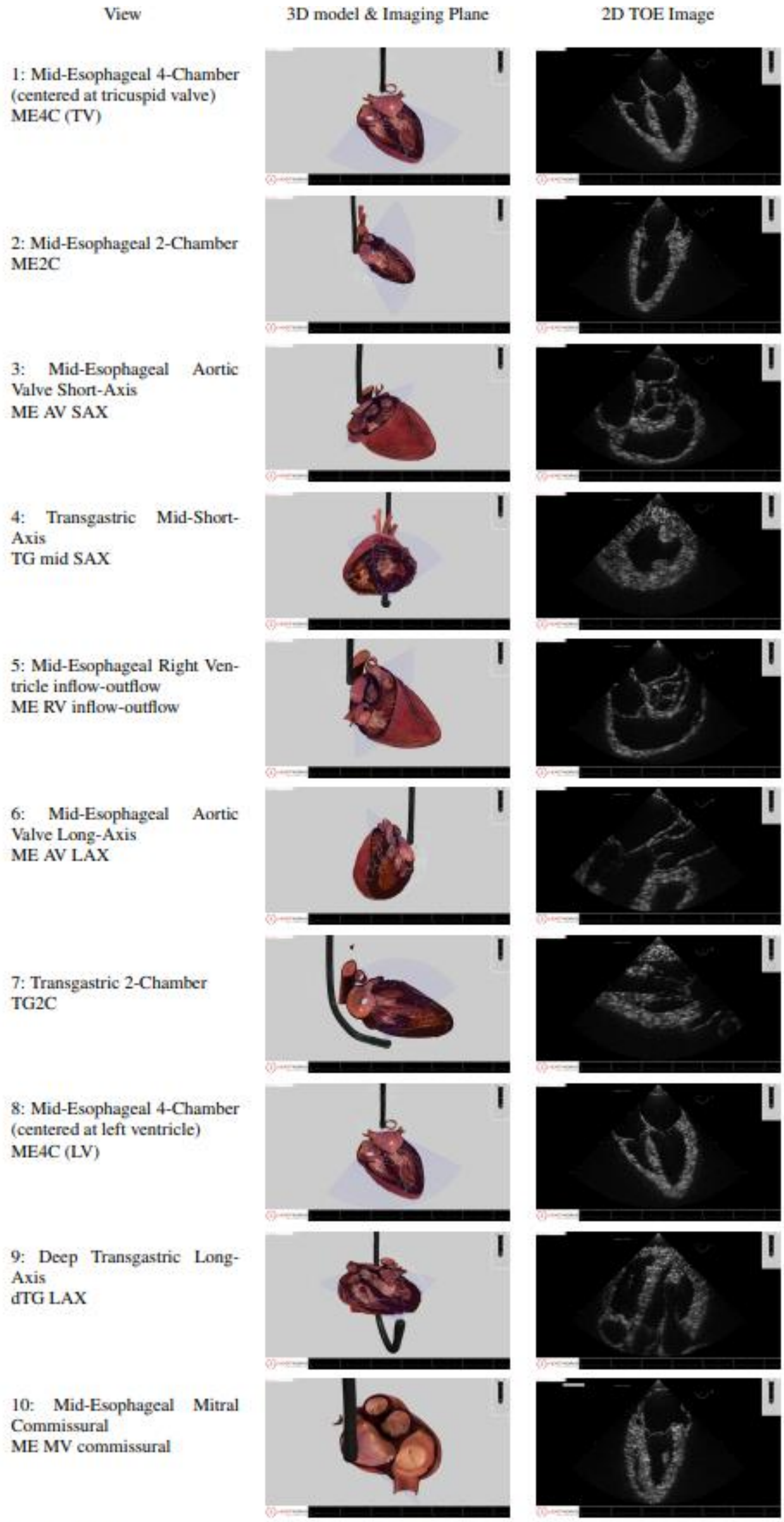


Fig. 2: The 10 suggested views used in the study, listed in acquisition sequence alongside correct position/orientation of the probe in the 3D model and the resulting 2D US image

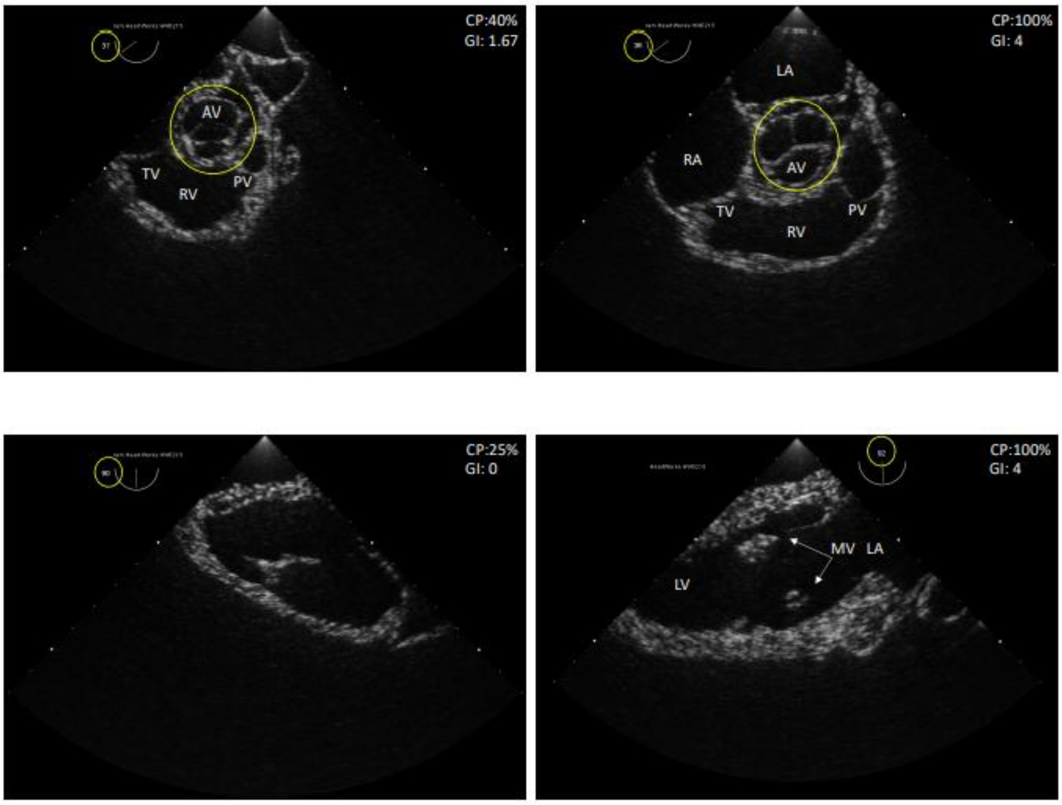


Fig. 3: Scoring examples for Views 3 and 7 from different participants with annotated the structures of importance. Left images are scored poorly whereas right images obtain excellent marks. Top row, View 3 - LA: left atrium, RA: right atrium, TV: tricuspid valve, RV: right ventricle, AV: aortic valve, PV: pulmonary valve, circle indicates visibility of AV cusps; Bottom row, View 7 - LV: left ventricle, LA: left atrium, MV: mitral valve and arrows showing leaflets.

High-resolution images from the 10 views were captured for analysis. A processing pipeline, shown in Fig. 4, includes a conditioning stage (opening and Otsu thresholding) to enhance the US image and eliminate specular highlights. Gaussian filtering is also applied to facilitate the segmentation of the heart structure in the US images performed using the Chan-Vese active contour algorithm.

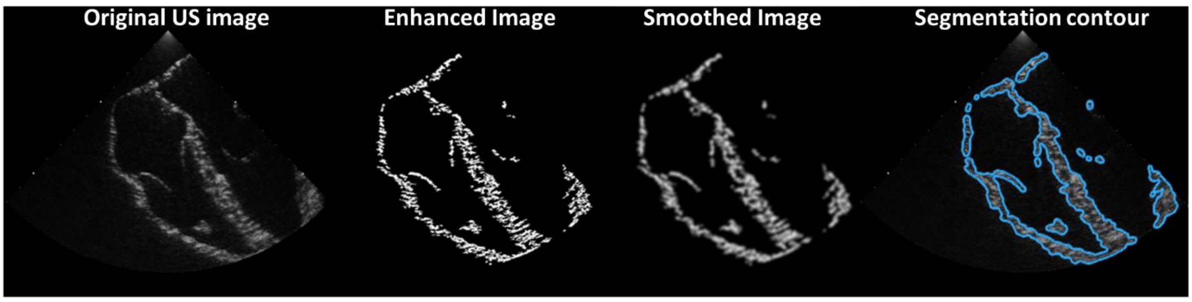


Fig. 4: The processing pipeline. The output is the segmentation contour of the heart's structure in the US image.

You can load the dataset stored in the cw2.mat file following this code:

```
import scipy.io as sio
import numpy as np
# Import the TOE image dataset from the cw2.mat file attached to Coursework 2 material on Moodle
cw2_data = sio.loadmat('path_to\cw2.mat') # Change to the path of the cw2.mat file
```

In total, the dataset contains 195 images (10 images/views per volunteer) because 5 images were not properly captured. The associated quality scores for **i)** criteria percentage score and **ii)** general impression score are also included. The file also has the 10 gold standard images.

The variables (numpy arrays) in cw2 are:

test_img: The 195 test images (20 participants x10 views array where each element is a 360 x 300 image)

gold_img: The 10 gold standard images (1standard x 10 views array where each element is a 360 x 300 image)

gen_impr: The general impression score (20 x 10 array where each element is a value 0-4)

crit_perc: The criteria percentage score (20 x 10 array where each element is a value 0-100)

For example: view1 of participant "1" can be accessed as: cw2_data['test_img'][0][0]

Participants 1-7 are the expert group and participants 8-20 are the novice group.

The missing images are in locations: [8][9], [12][7], [13][9], [14][0], [15][3] and these are empty arrays. The corresponding score values for general impression and criteria percentage are "-1". These must be ignored in your analysis.

NOTE: For your analysis in this coursework, you can use the "scikit-image" python libraries. It can be installed in your "sds" anaconda environment with the following command: "pip install scikit-image". You can then use it in your code after you declare with "import skimage".

You can also use the opencv library (pip install opencv-python) and declare it with "import cv2". You can then use the cv2.findTransformECC method to perform the image alignment. These will be useful to derive the parameters in Question 4.

Follow this link (https://github.com/TheJark/Image-Matching/blob/master/image_align2.py) as reference on the usage and consider the parameters given below:

```
# Define the motion model
warp_mode = cv2.MOTION_EUCLIDEAN

# Define the warp matrix
warp_matrix = np.eye(2, 3, dtype=np.float32)

# Specify the number of iterations.
number_of_iterations = 500

# Specify the threshold of the increment
# in the correlation coefficient between two iterations
termination_eps = 1e-10

# To otbain the aligned test image (line 46 of the link) use:
test_aligned = cv2.warpAffine(test, warp_matrix, (sz[1],sz[0]),
flags=cv2.INTER_LINEAR + cv2.WARP_INVERSE_MAP)
```

Question 1 [20 marks]

Considering the two different image quality scores (general impression, criteria percentage):

i) Calculate the Pearson correlation coefficient for the two scores for each view. Identify for which view the two scores are in higher agreement. (5 marks)

ii) For each view, perform linear regression analysis for the two image quality scores (use general impression as dependent variable and criteria percentage as independent variable). Compute the RMSE and R^2 scores and comment on the performance of your regression. (10 marks)

iii) For each view, plot the true vs estimated values for the three best performing views and comment on the model performance for two different criteria percentage score ranges (0-2, 2-4). (5 marks)

Question 2 [20 marks]

Similarity metrics

Given two images with similar dimensions: I_a, I_b

$$SSI = SSI_l \cdot SSI_c \cdot SSI_S \quad \textbf{(1)}$$

$$SSI_l = \frac{2\mu_a\mu_b + C_l}{\mu_a^2 + \mu_b{}^2 + C_l}, SSI_c = \frac{2\sigma_a\sigma_b + C_c}{\sigma_a^2 + \sigma_b{}^2 + C_c}, SSI_s = \frac{\sigma_{ab} + C_s}{\sigma_a\sigma_b + C_s}$$

μ_x: mean, σ_x : St. deviation, σ_{xy}: covariance,

C_l, C_c, C_s: stabilising constants

$$E_x = -\sum_x p_x(x) \log p_x(x) : x = a, b \quad \text{Image entropy}$$

$$E_{a,b} = -\sum_{a,b} p_{ab}(a,b) \log p_{ab}(a,b) \quad \text{Joint entropy} \quad \textbf{(2)}$$

$$MI = E_a + E_b - E_{ab}$$

$$|I_a, I_b| = \frac{I_a \cdot I_b}{||I_a||||I_b||} \quad \text{Cosine similarity} \quad \textbf{(3)}$$

$I_a \cdot I_b$ is dot product of the image I_a and image I_b

$$||I_a|| = \sqrt{I_a \cdot I_a}$$

We are interested in evaluating the content similarity of the test images against the gold standard ones. To perform this, we will use two key statistical metrics developed to compare the content between two images. The structural similarity index (ssi), proposed by Wang et. al. [3], which separately compares three components; luminance (ssi_l) represented by the mean intensity, contrast (ssi_c) represented by the standard deviation of the intensity and structure (ssi_s). The mutual information (MI), an entropy-based index that in essence measures the amount of information that one image contains about the other [4], thus representing content similarity. We will also calculate the cosine similarity, one of the most used similarity measures, which is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. In our case these vectors are the two images. These metrics are defined according to the following equations: structural similarity index (SSI) - Eq.1; mutual information value (MI) - Eq.2; cosine similarity (CS) -Eq3. To compute the cosine similarity, you will need to reshape your images to vectors (also known as flattening and can be easily done in python). Please use as reference the code written in this article.

i) Calculate the SSI, MI and CS values for each test image against the gold standard ones (use the scikit-imag, scikit-learn, scipy libraries). Identify and report the top three test images (which participant) for each view that have the most similar content to the gold standard ones according to their SSI, MI and CS values. (10 marks)

ii) For each view develop a hypothesis and perform a statistical test to evaluate the differences between the expert and novice groups in terms of SSI, MI and CS. Discuss your results in terms of significance. Which similarity metric better shows the differences between expert and novice surgeons? (10 marks)

Question 3 [30 marks]

Considering the extracted SSI, MI, and CS values:

i) Calculate the correlation coefficient for SSI and MI, SSI and CS, and MI and CS for each view. Identify for which view, in each pair above, the two parameters are in higher agreement. (5 marks)

ii) Perform polynomial regression using a 7th degree order polynomial for the SSI, MI, and CS against both the criteria percentage and general impression for each view (SSI/MI/CS – independent, manual scores – dependent). In case this leads to overfitting use a regularization method (LASSO, RIDGE or Elastic Net regression) and identify the optimal degree of the polynomial to avoid overfitting. Justify the selection for the regularization method. Calculate and list the RMSE and R2 scores of your regularized regression. Identify the three views for which the regularized regression performs better. Plot the regression output only for the three best performing cases. Consider regression coefficients smaller than 0.01 as not contributing to your regression. (15 marks)

iii) Perform linear regression using Gaussian basis for the SSI against the general impression score for each view (SSI – independent, general impression – dependent). Decide the order of the gaussian basis, so that the regression model will not underfit/overfit (you can also use a regularization method to identify that). Calculate and list the RMSE and R2 scores of your regression. Identify the three views for which the regularized regression performs better. Plot the regression output only for the three best performing cases. Use the code provided in the tutorial 4 on Moodle to develop your regression model. (10 marks)

Question 4 [30 marks]

As we can see from Fig.3 a good quality US image means that the heart structures are correctly placed in the centre of the US scan. To evaluate this, we will calculate the degree of misalignment of the test images against the gold standard ones. To achieve this, you will use the ECC algorithm [5], to derive a rigid transformation that represents the misalignment and assess the differences in position and orientation of the captured US image against the gold standard one (refer to page 7 for instructions). The 2d displacement and angle of rotation can be extracted from the resultant alignment matrix as follows.

Consider a 2x3 matrix M to describe the rigid transformation:

$$M = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \end{bmatrix}$$

The translation (in pixel units) is :

$M = \begin{bmatrix} m_{13} \\ m_{23} \end{bmatrix} = \begin{bmatrix} m_x \\ m_y \end{bmatrix}$ so the total displacement is: $\sqrt{m_x^2 + m_y^2}$

And the abstract rotation angle in radians is: $\text{atan}\,(m_{21}, m_{11})$

i) Calculate and list the rotation (in degrees) and translation values (in pixel units) of the rigid transformation (use the opencv library) between the test and gold standard images (10 marks)

ii) For each view develop a hypothesis and perform a test to evaluate the differences between the expert and novice groups in terms of the values of rotation and translation. List and discuss your results in terms of significance. (10 marks)

iii) Perform linear regression for the rotation and translation against both the criteria percentage and general impression (rotation/translation – independent, manual scores – dependent) for each view. Calculate and list the RMSE and R^2 scores of your regression. Identify the three views for which the linear regression performs better in every combination of independent/dependent variables. Plot the regression output only for the three best performing cases. (10 marks)

References

[1] E. Mazomenos et. al. “Automated Performance Assessment in Transoesophageal Echocardiography with Convolutional Neural Networks”, MICCAI 2018

[2] M. Cheitlin et. al. “ACC/AHA guidelines for the clinical application of echocardiography: executive summary. A report of the American College of Cardiology/American Heart Association Task Force on practice guidelines (Committee on Clinical Application of Echocardiography). Developed in collaboration with the American Society of Echocardiography”. JACC 1997.

[3] Z. Wang et. al. “Image quality assessment: from error visibility to structural similarity”. IEEE TIP 2004.

[4] F. Maes et. al. “Multimodality image registration by maximization of mutual information”. IEEE TMI 1997.

[5] G. Evangelidis and E. Psarakis. “Parametric image alignment using enhanced correlation coefficient maximization". IEEE TPAMI 2008.