Final Year Project (KIE4002)

Progress Report 1

Automated Facial Landmark Detection in Medical Images for Patient-Image Registration in Cranial Surgeries (MM5)

by

Abdolraouf Rahmani

KIE170720

Supervisor :

Prof. Mahmoud Moghavvemi

Department of Electrical Engineering

Faculty of Engineering

University of Malaya

2020

Abstract:

Medical image analysis is a critical element in successful diagnostics and treatments. Registration is one such critical image analysis. In particular for surgeries, an accurate registration will allow surgeons to make minimal incisions which reduces surgery duration and patient recovery time. Due to presence of large number of facial landmarks, an intraoperative facial image of patient can be registered onto preoperative image. Traditionally radiologists identify such facial landmarks manually which is time consuming and prone to human error. To improve the registration workflows automatic facial landmark detection is proposed. Automating facial landmark detection is not new but there are less studies on finding landmarks on medical images (such as MRI and CT). This FYP project aims to evaluate a functional algorithm for this application.

Table of Contents

1 Chapter 1: Introduction & Motivation

This chapter will highlight the importance of automation of image analysis process in health care (section 1.1) and then further discuss the importance of registration using facial landmark detection specially for cranial surgery (section 1.2).

1.1 Automated Image Analysis in Healthcare

Images are the largest source of data in healthcare and so require intelligence to be analyzed. In absence of medical analysis software, clinicians would rely on overworked radiologists to interpret the data. Such interpretation is prone to negligence, oversight, and mistake.

The benefits of algorithm in medical imaging will benefit many stakeholders such as radiologists, non-radiologists, clinical patients, and medical institutions among many others.

Radiologists: Unfortunately there is a limit to number of radiologists to cope with the growing number of tasks in analysis of MRI, CT, PET and ultrasound data. For instance conventional method of image registration was to have it be done manually by such individuals. Automating such process will ease their burden by either only require their attention on flagged data or providing supplementary information to ease task accomplishment. For instance tumor detection algorithms will provide the likelihood of a tumor being either benign or malignant and segmentation of tumor, which helps doctors to focus their work on patients that require immediate attention, and support their diagnostic decision respectively.

Non-Radiologists: Imaging algorithms have the potential to increase accessibility to radiology by enabling non-radiologists such as in under-served areas or paramedics to utilize this expertise, perhaps on their mobile devices.

Patients: with medical imaging assistance from machine algorithms, patients will not only get a more reliable treatment or diagnosis, they will also have to wait a lot less for such process to finish . For instance accurate registration algorithms will help smaller incisions for tumor resection surgery., which will reduce the inpatient duration for recovery.

Medical institutions: places like hospital will be able to increase their revenue. This can be done by reducing manpower, reducing inpatient periods so new patients can be accommodated as well as providing high quality diagnosis, and treatments.

As was highlighted one of the image analysis algorithms is registration of different image data together-different in terms of image modality, or time an image was taken- to increase information density. Since This paper focuses to improve the registration algorithm the background of registration in medical imaging is provided next.

1.2 Image Registration in Health Care

Registration is an image analysis technique not only unique to medical field. Registration is a process of alignment and transformation of two images of the same subject. For instance registration of two images of the same part of the body is carried out to determine the correspondence between them. The images will be different depending on the modality of image-CT, MRI, or ultrasound, the time of capture, or images belonging to different patients. Furthermore registration is carried out when corresponding digital atlas data with acquired image. The variations between images may be due to different imaging methods, artifacts-natural and unnatural misrepresentation in images, deformable subject and natural variations between patients.

Registration of intraoperative and preoperative patient images is very important during surgery as the result of registration is used to make critical decsions. In cranial registration for surgery, surface image of face, obtained from RGB-D (red green blue-depth) camera, for intraoperative image of patient and MRI image for preoperative image can be used. Using RGB-D for intraoperative image is important due to rapid image acquisition. Furthermore in recent years there

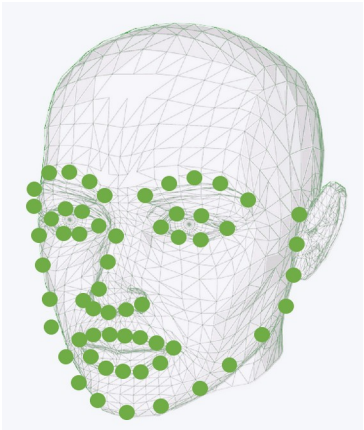


Figure 2: 3D face model annotated by facial landmarks. Number of landmarks is dependent on the rules used by the feature extractor.

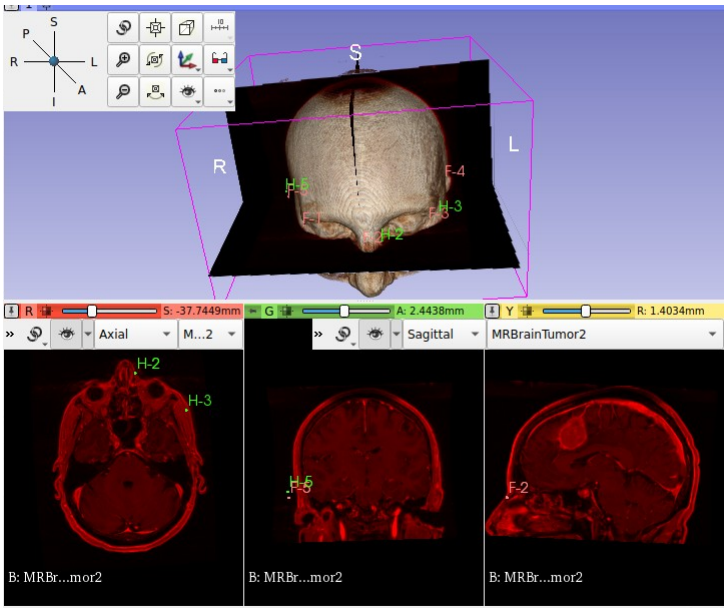


Figure 1: Facial landmark being used for registration. The landmarks labeled as Hx, where x is number from 1 to 5, denote landmarks of first cranial MRI (not shown here). And the landmarks Fx where x is number from 1 to 5, denote landmarks from second cranial image. The cranial image shown is the result of registration of the two aforementioned images.

has been many reliable and accurate depth cameras with reasonable prices such as The Intel® RealSense™ depth camera D435. A major challenge in registration of aforementioned image modes is the difficulty to obtain features as basis for registration. A normal method of feature extraction is to have an expert identify the facial landmarks manually. This process is time consuming and tedious. To create synergy between the operator and the image-guided system, automated methods for extraction of these landmarks have been developed.

1.2.1 Problem Statement

Conventional methods for landmark detection is to have radiologists manually annotate them. This is time consuming and prone to human error. Such drawbacks become critical in surgeries. There are many facial landmark detection algorithms, but very few of them are applied to medical imaging and their unique characteristics.

1.2.2 Objectives

The objectives for FYP are as given below:

1. Compare neural network based landmark detection over traditional methods.
2. Evaluate a technique for segmentation of face from the rest of head model.
3. Obtain reliable model for facial landmark detection of medical images

As of end of first semester objective one has been accomplished to a significant extent, and objective 2 has been done using manual methods, so what remains is automating the process. Due to the massive number of techniques and combination of techniques used in facial landmark detection only the significant techniques and algorithms were used to accomplish objective one.

2 Chapter 2: Literature Review

This chapter includes theoretical foundations as well as critical review of recent publications. There are two sub-chapters, (1) Theoretical foundation and (2) Critical Review of Previous Work.

2.1 Theoretical Foundation

2.1.1 Segmentation

To achieve objective 2- "Evaluate a technique for segmentation of face from the rest of head model", basic understanding of segmentation is required.

There are numerous diverse proposed segmentation techniques being used and researched on, but a common method for all application types is still unavailable. In general, segmentation partitions images into different groups based on shared properties. Hence each group will have a homogeneous property. Common properties used for segmentation are color, intensity, texture and reactivity. Medical imaging dominantly makes use of intensity based segmentation.

This section introduces categorization based on abundance of usage. Figure below lists common examples under categories of obsolete, ancient and recent techniques for segmentation.

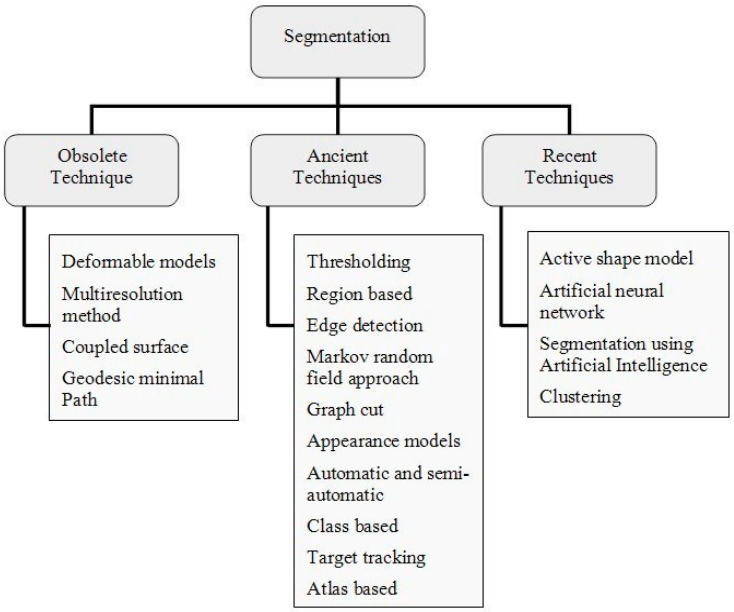


Figure 3: Categorization of Segmentation based on period of usage.

Availability of techniques shown in figure above is dependent on the software used. Nevertheless there are techniques that are found across all software and are useful. Such technique is thresholding.

Thresholding is a simple yet powerful means of segmentation, which is categorized under local and global thresholding . Global thresholding consists of setting an intensity value (threshold) such that all voxels having intensity value below the threshold belong to one phase, the remainder belong to the other. Unlike the global thresholding technique, local adaptive thresholding chooses different threshold values for every pixel in the image based on an analysis of its neighboring pixels. Reliability of thresholding lies on the value of threshold which is determined visually by operator depending on the shape of histogram, or automatically by optimization techniques to compensate for background noise, Otsu's method, texture analysis using gray-level concurrence matrix, and Kapur's posterior maximum entropy method. (Sheeba & Manikandan, 2014) .

2.1.2 Mathematical model: Regression

Inorder for computer to make a prediction based on input they must you models. Models are mathematical representation of knowledge. For example humans have the knowledge to locate the tip of nose. To allow a computer to carry out the same complex task it must have relevant knowledge. One this knowledge is learn via training of model, computer can infer, predict, where the nose tip is located. This learning of machines to carryout complex tasks is called machine learning (ML). There are many different models used for different applications, here the regression model is introduced as it is predominantly being used for image processing.

Machine learning tasks can be gathered into the four following categories:

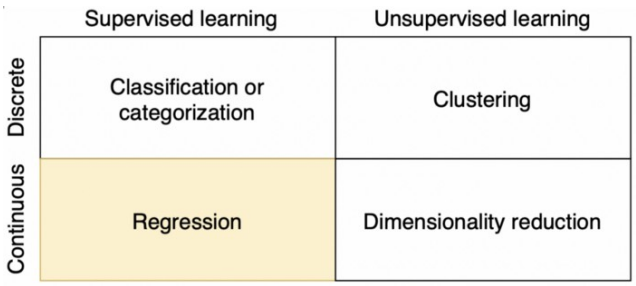


Figure 4: Different types of models based on output type and data form used for model training. Regression models are trained on labeled datasets and are used to produce a continuous output.

Regression is defined as: "*In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features')*". Regression analysis is primarily used for prediction and inferring relationships between

independent and dependent variable. Unlike classification the output of regression models is numerical continuous as opposed to discrete.

Regression models are evaluated using root mean squared error or other types of loss functions. The loss function is used to change the regression parameters during training until the loss is minimum. Every regression technique has some assumptions attached to it which we need to meet before running analysis. These techniques differ in terms of type of dependent and independent variables and distribution.

The most common model in regression analysis is linear regression. This model finds the relationship between the independent and dependent variables by fitting a linear equation. The most common method for fitting this regression line is using least-squares, which calculates the best-fitting line that minimizes the sum of the squares of the vertical deviations from each data point to the line. There are many more regression models as some are listed as below:

1. Polynomial Regression
2. Support Vector Regression
3. Decision Tree Regression
4. Random Forest Regression
5. Ridge Regression
6. Lasso Regression
7. Logistic Regression

Individual regressors and ensemble of regressors are a useful method of predicting the facial landmark locations. Further detail is covered under section 4.1.1.4.

2.1.3 Landmark Estimation errors

The simplest comparison is a root mean squared error (RMSE) assessment; where the average distance between each of the N predicted landmarks (x^p_i, y^p_i, z^p_i) and the corresponding 'ground truth' (x^t_i, y^t_i, z^t_i) is calculated on a per landmark basis. Landmarks that are poorly predicted will be positioned far their corresponding ground truth locations and thus contribute to increasing the RMSE value. Often the root mean squared error is normalised by the distance between two specific 'ground truth' points (NMRSE) such as the left (x^t_{le}, y^t_{le}, z^t_{le}) and the right (x^t_{re}, y^t_{re}, z^t_{re}) outer corners of the eyes d_{norm} (see Eq.) [10] to allow a fair comparison between faces of different sizes

$$RMSE = \frac{1}{N} \sum_{i=1}^{N} \sqrt{(x_i^p - x_i^t)^2 + (y_i^p - y_i^t)^2}, \quad (1)$$

$$NRMSE = \frac{1}{N} \frac{\sum_{i=1}^{N} \sqrt{(x_i^p - x_i^t)^2 + (y_i^p - y_i^t)^2}}{d_{norm}}, \quad (2)$$

$$d_{norm} = \sqrt{(x_{\text{le}}^t - x_{\text{re}}^t)^2 + (y_{\text{le}}^t - y_{\text{re}}^t)^2} \quad (3)$$

When comparing the performance of different landmarking algorithms against the same dataset, the average RMSE or NRMSE value over the number of samples in the dataset (K) may simply be reported. A more detailed summary of the model performance can be provided using the cumulative error distribution (CED), which plots the cumulative NRMSE against the proportion of images with an NRMSE of less than or equal to a particular value. (Johnston & Chazal, 2018)

A less frequently reported metric is the landmark detection rate, i.e. the proportion of the N landmarks from the K images, correctly identified by the system. A landmark is correctly identified if its position is less than a defined Euclidean distance from the ‘ground truth’. Similarly to the mean squared error calculations, landmark detection rate can also occur on a per-image, per-landmark, and overall average basis.

2.2 Critical Review of Previous Works

Before studying and reviewing recent publications, it is important to define the parameters of landmarking for this project. On one hand, emerging applications require that the landmarking algorithms run in real-time while operating with the computational power of an embedded system, such as intelligent cameras. On the other hand, some applications require increasingly more robust algorithms against a variety of confounding factors such as out-of-plane poses, occlusions, illumination effects and expressions. The details of these confounding factors that compromise the performance of facial landmark detection for this project is determined as below:

- **Variability:** Landmark appearances differ due to intrinsic factors such as face variability between individuals, but also due to extrinsic factors such as partial occlusion, illumination, expression, pose and camera resolution. Given MRI images are the data analyzed, this project requires relatively high intrinsic variability but very low extrinsic variability.
- **Number of landmarks and their accuracy requirements:** minimum number of landmarks is to be 11 (6 eyes, 3 nose, 4 mouth). It has been recorded by (Çeliktutan et al., 2013), with increase in 3 to 68 landmarks, there is 50% accuracy improvement, hence higher landmark numbers are favorable. Given that landmarks are to be used for registration, they are to be accurate within +-2mm to expert annotators (Liu et al., 2017).

Relevant papers can be divided into two groups based on method of approach to obtain 3D facial landmarks. Firstly is to analyze the 3D model and identify landmarks on it. Secondly is to convert the model into a representative 2D image and identify the landmarks. This chapter consists of two sub-chapters, (1) 2D representative Method, and (2) 3D Methods. Each sub-chapter will involve summation of relevant points from the papers and then their review in accordance to the project requirements.

Search Expression: (mri OR medical image OR ct OR 3d OR 3-d OR “three dimensional”) (facial OR face) (landmark OR marker OR feature OR fiducial OR keypoint) (detect OR localization OR localize OR formalization OR identification OR identify)

2.2.1 2D Representative Method

The abundant of research is on facial landmark detection of 2D data. Such algorithms can be used as an inspiration to develop it's 3D counter part, or The 3D image can be projected into it's 2D representative to locate the 3 dimensional landmark.

2.2.1.1 Paper 1

Literature Title: Real-time facial feature detection using conditional regression forests (Dantone et al., 2012)

Paper Summary:

This paper makes use of conditional regression forests to develop real time landmark detection with low quality images. They justify use of conditional regression forest over regression forest by point out the conditional element will introduce more generality for the model and so it makes it more robust for deformed faces. Regression forests in general are responsible for making relationship between image patches and facial landmark location , with certain probability. Conditional regression forest are more general as this relationship is made conditional to global properties, such as head pose.

First part of their training workflow of regression forests consisted of obtaining annotated images. These images are divided into exclusive sets and are used to train different trees. Each image is made to produce random number of patches that contains appearances and facial landmark offsets. Appearances are the patch intensity, normalized intensity and filtered intensities. Facial landmark offsets are the displacement vectors from that patch to landmarks. The global patch set is then split into many two subsets, to give many separations made up of these two subsets. So subset P_L and P_R are different for a given separation,φ. One of the separations ,φ, is selected and used to evaluate the relationship of it's patches with it's neighbors. Using an evaluation function called Information Gain (IG), either a leaf is created or different separations are selected to be evaluated. A leaf is created if the IG is below a threshold or maximum predefined set is reached. This is iterated to maximize the evaluation function which would reduce uncertainty for a separation. The uncertainty indicates the probability of a facial landmark to a given patch set. This model training workflow is that of any regression forests.

Both conditional and non conditional regression forests develop the same trees. Difference between the models is how the trees are analyzed. For unconditional regression forests the probability of a

forest is obtained by averaging the probability for a patch ending in the leaf of a single tree. conditional uses a more complicated method that is dependent on a global property.

Review

Accuracy of landmarks deviate from 71 to 92 percent. Such deviation is not reliable for use in medical image registration. This paper demonstrates the benefit of using random forest to provide a more generalized model. This was deemed insufficient and so a new feature was added to the normal regression forest model development workflow. This shows the flexibility of random forests as a regression model.

2.2.1.2 Paper 2

Literature Title: One Millisecond Face Alignment with an Ensemble of Regression Trees (Kazemi & Sullivan, 2014)

Paper Summary:

Their innovation is the use of 2 elements from past researches resulting in a cascade of high capacity regression functions learnt via gradient boosting.

Instead of regressing shape parameters based on the features of raw image, the normalized image obtained from an initial estimate is used. This is their first novel element. This process is repeated several times until there is a convergence of the shape parameters.

Second element is the assumption that the predicted shape lies in a linear subspace. Based on this assumption the number of potential shapes considered for inference/prediction is reduced. This helps avoiding the local optimas. To implement this a regression model that predicts shapes in a linear space is used. To produce such regression models gradient boosting is used.

Regressors in arranged in cascade which means that each regressor in the cascade predicts an update vector from the image and the result of the update vector from the previous regressor. Furthermore the cascade can be trained on images with missing labels. There are weak and strong regressors in cascade.

Through experiments it was found that with increase in number of strong regressors in cascade the performance was improved with little increase in computation. Using 500 weak regressors and 10 strong regressors the performance was compared to with active shape model. Cascaded regression had an error of 0.049 while active shape had error of 0.111. They further point out that major components of algorithms treat different target dimensions as independent variables.

Review

For comparison they compared different ensembles of regression trees as well as active shape models. The active shape model they used dates back to 2008 and so doesn't make for a reliable comparison. In terms of error as well as speed they provide good performance, furthermore since major components of their algorithms treats different target dimensions as independent variables. It means there is room for grow with cascaded regressors.

2.2.1.3 Paper 3

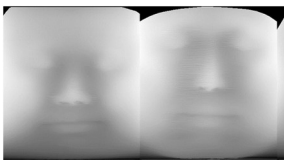
Literature Title: 3D Facial Landmark Detection Using Deep Convolutional Neural Networks (Terada et al., 2018)

Paper Summary:

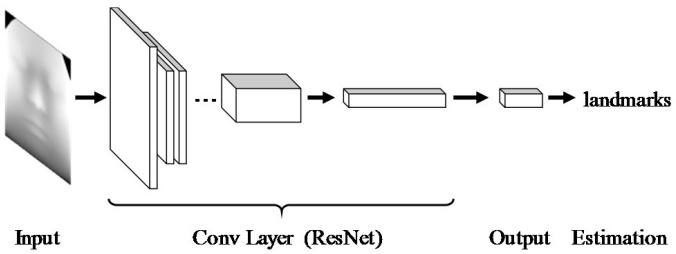
In order to project 3D image this paper uses cylindrical projection. Given that the point cloud of face surface is given by Cartesian coordinates, points are converted into cylindrical coordinate system as follows:

$$(\rho, \theta, z) = \begin{cases} \rho = \sqrt{x^2 + y^2} \\ \theta = \tan^{-1}\frac{y}{x} \\ z = z \end{cases},$$

A 2D representative is obtained by getting p(θ, z) where the pixel value becomes the p value. Furthermore the projected image is coped to only give the frontal face image.



input image is a cylindrical projection of 3D face model. This paper makes use of regression function whose implementation uses CNN. The CNN network architecture used was ResNet:



The loss function used, mean square error, to assess the performance of model. Loss function is given as below :

$$L_r = \frac{1}{N}\sum_{k=1}^{N}(p_i - \hat{p_i})^2.$$

ResNet18 and ResNet34 were used with and without data augmentation. Augmentation dataset used translation and scaling to increase the variability and quantity of data.

After obtaining the landmark, the coordinates were inverted into Cartesian form using:

$$(x,y,z) = \begin{cases} x = \rho\cos\theta \\ y = \rho\sin\theta \\ z = z \end{cases}.$$

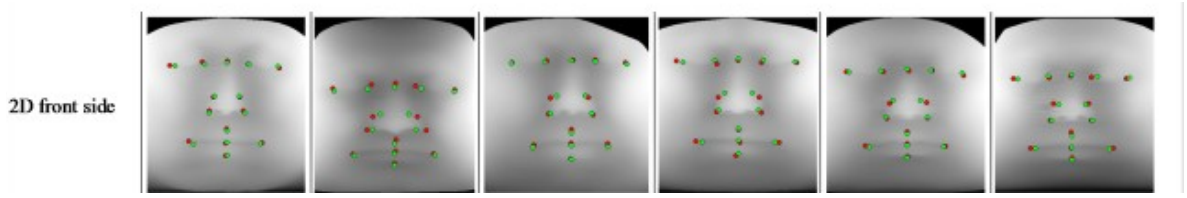
The 2D image generated for identification model was sampled at 360 horizontal and 400 vertical to give 360 x 400 2D image.

Review:

The paper doesn't mention about the location of the origin or the scaling and equation of transfer function p. This choice will effect the resolution of the obtained image. Best origin location is assumed to be located at same level as tip of the nose and lying on the vertical rotation axis of head. This would ensure good pixel value deviation.

Furthermore they mention that the ears and the neck should be ignored, which could imply that image is to be segmented using face detection algorithms.

The image that they have generated is not very clear and has little information on specific edges, shade of facial part, wrinkle and skin color. There is lack of information on best scales for p(θ, z) which would affect 2D image resolution as well as contrast. Furthermore it has error of 11 pixels using ResNet34 with data augmentation. The error given is the mean error for 14 points and it is hypothesized the landmark numbers can not be increase as the error could become more due to lack of image clarity.



It is not possible to evaluate this technique without distance errors in 3D as well as registration error, eventhough it unlikely that the error is less than +-2mm.The use of cylindrical projection is a very useful method as it would allow for 2D representative image generation that doesn't require the landmarking algorithm to carryout significant geometric normalization. Reducing burden on normalization could mean more accuracy as the main purpose of using 3D images for facial

landmark detection is that 2D images have more geometric variations, among other variations. There is no justification for use of CNN model in this paper.

2.2.2 3D analysis

facial landmark detection on 3D data is considered to be more accurate but to large amount of information in such data normalizing it is a difficult task. 2 papers carrying this method are reviewed in the following sections.

Literature Title: Facial landmark automatic identification from three dimensional (3D) data by using Hidden Markov Model (HMM)(Liu et al., 2017)

Given in paper:

Generally, there are two major issues regarding the anatomical landmark identification, i.e., local feature extraction and identification algorithm. They use spin image (SI) to extract the local features and Hidden Markov Model (HMM) to implement the landmark identification. 11 HMMs were trained and evaluated in this study.

RBFANN performance is lower than HMM but it’s doesn't mean all ANN are bad. There are many ANN techniques such as BPANN and ANNs with heuristic algorithsm such as GA and SAA that improves inference effect.

3 reasons why ANN may not be as good here:

- take longer specially when network size is large
- rarely it is used for dynamic patters as it has poor descriptive ability for this case.
- Easily confined in local optimums which leads to negligence of optimum value.

Data and Sampling: slice interval of 5mm. Sample size 120 with 23,000 points per sample.

Method used: Spin image representation of orientation points are used to characterize the local features. Local features with coordinate given by x are converted to cylindrical coordinate (alpha, beta) by:

$$\alpha = \sqrt{\| x - p \|^2 - (n \cdot (x - p))^2} \quad (1)$$

$$\beta = n \cdot (x - p) \quad (2)$$

Only values of beta and alpha that meet a certain criteria are placed in the spin image. The spin image of given orientation point is very redundant and due this dimensional reduction using

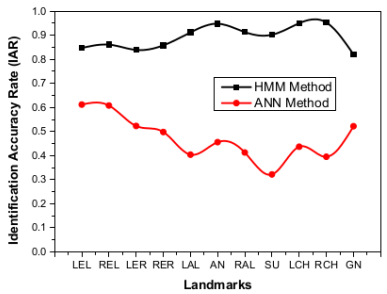
$$f_m = \sum_{k=1}^{n-1} x_k \cos\left[\frac{\pi}{n} m \left(k + \frac{1}{2}\right)\right]$$

Discrete Cosine Transform type II is used.

This reduces the dimension from a 2D matrix of alpha and beta to a vector. The geometrical features of each landmark are characterized by their projection after dimension reduction.

HMM models with 3 to 8 states, and observable states that provide the SI of the landmark is built to predict each landmark. Baum-Welch algorithm was used to train the landmark HMM. The discrete HMM is iterated till following a criteria is met. The goal of probabilistic inference over HMMs is to find optimal matching between random landmarks and the trained HMMs, which has the maximum matching probability. The SI s that provide the maximum probability and meet the predefined threshold are identified as landmarks.

Results: The maximum landmark recognition rate is 95.9%. with +- 1.6mm vertical and 3mm horizontal error. Effect of SI parameters ,such as Bin Size and Support Angle, on IAR (identification accuracy rate) were empirically studied. It is better than ANNs in accuracy, robustness and adaptability.



Evaluation: Three hierarchy experiment is used to test the validity and reliability. The three hierarchy preliminary experiment consisted of single palpating on nasal part by single participant,

repetitive palpating on nasal part by single participant and single palpating on nasal part by a population of participants. The procedure is well defined for all 3 experiments.

Review: They claim the reason their error is somewhat high because the sampling interval was no more than 1.6mm due to data having slice interval of 5mm as well as due to hole filling and noise reduction. And so if a different sampling method is used the level of error may reduce. There is no citation of ANN application in facial landmark detection and so the paper feels biased and requires more investigation. Overall they pitch a promising method. This paper makes use of local feature extraction so what of global feature extraction? Why is that not possible?

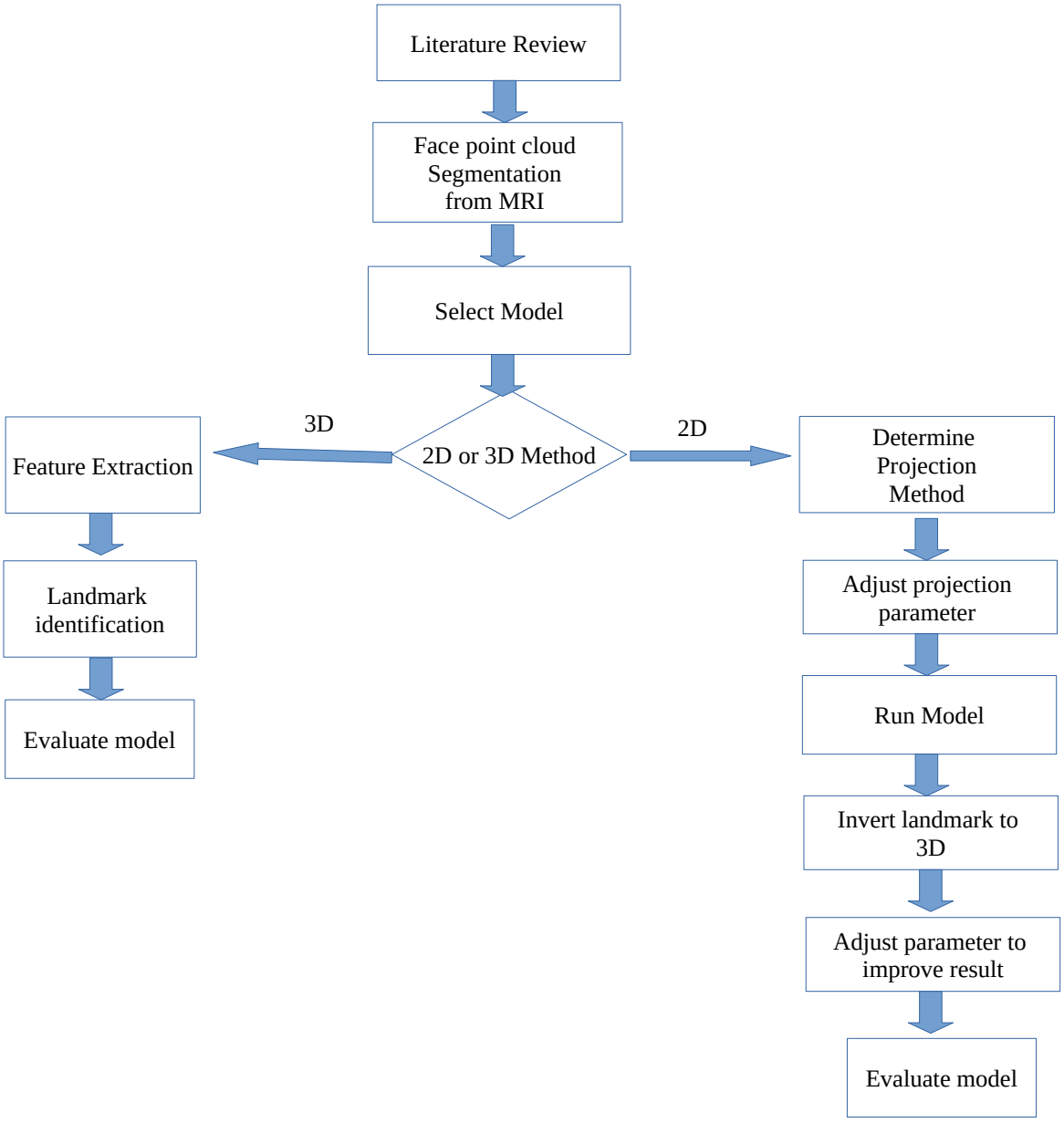
The dataset used is only for Chinese soldiers in a given range, so the reliability of model in general has not been assessed.

2.2.3 Literature Review Overview

3D-to-2D methods make use of projection techniques such as orthogonal, perspective, and most used technique: cylindrical projections. The 2D image is then used to obtain the facial landmarks whose methods can be divided into three categories: generative methods, discriminative methods and methods combining the two, producing statistical methods. On the other hand 3D facial landmark detection is made up of two components, feature extraction and landmark identification. There can be many combinations to make up the 3D facial landmark detection, which makes selecting a suitable algorithm very difficult. In addition to that abundance of research is on 2D facial landmark detection and less on 3D facial landmark detection.

3 Chapter 3: Methodology

This chapter discusses the methodology to construct an automatic facial landmarking system with performance (specially accuracy) comparable to that of an expert human annotator, using MRI images. To do so the work flow below is carried out to select to best model for application:



Models

There are many models being implemented for 3D facial landmark detection. To determine the right one models need to be compared. 2 models for 2D and 2 for 3D methods is selected and they will be compared. One of the 2D models will be based on cascaded regression, and the second based on neural networks. The models for 3D methods will be selected during the school break.

Dataset

Some models require to be trained on labeled datasets, so a dataset must be selected for such models. Database of MRI scan with annotated facial landmarks are scarce and alternative to that is use of BrainWeb: Simulated Brain Database (Cocosco et al., 1997). This is pre-computed simulated brain database (SBD), the parameter settings are fixed to 3 modalities, 5 slice thicknesses, 6 levels of noise, and 3 levels of intensity non-uniformity. Despite it's very early initial release it is still being used, for example (Qi et al., 2020). The subset of database used will be the normal brain database with following characteristics:

Characteristics	Values/types		
Modality (pulse sequence)	T1	T2	PD
Slice thickness/mm	1	3	5
Noise (relative to the brightest tissue)/ %	0	1	3
Intensity non-uniformity ("RF")/ %	0	20	40

The dataset covers large variation of data to ensure results are of the general cases. The software used will be 3DSlcier and using it's ITK library the MINC file format will be loaded. This database unlike others is completely open source.

There are other relevant databases that were considered:

Name	Description	Data	Need to register	Ref.
Neuromorphometrics.com	Manually labeled MRI Brain Scans	T1-weighted MRI, labeled volumes	No	
SchizConnect	SchizConnect is an open, public search-and-download virtual database for schizophrenia neuroimaging (MRI) images and related data.	Structural, Diffusion and Functional MRI datasets, cognitive and clinical assessments	yes	(Wang et al., 2016)
Ultrahigh resolution T1-weighted whole brain MR dataset	T1-weighted MR data acquired using prospective motion correction at an ultrahigh isotropic resolution of 250 μm.	Structural MRI dataset including scanner's raw to processed data	No	(Lüsebrink et al., 2017)

Face Point Cloud Segmentation

For automated facial landmarking we must detect and extract the face from the image and discard irrelevant information such as the background. This face detection process forms the first stage of an automated landmarking system and is critical for overall performance. The system must accurately identify and locate the face within the image, given variations in lighting, pose, expression, and face appearance. To do so the process involves firstly skin segmentation and then clipping skin to only get the ROI, which is the face point cloud starting from the end of ear to the nose tip. All MRI data are stored to give an image represented in same manner in the coordinate system, which makes the task relatively simple.

Model Training and Evaluation

The selected model is trained on dataset to give best predictions. For evaluation of the model normalized root mean squared error (NRMSE) will be used to compare the performance of different landmarking methodologies. The formula used is given in theoretical background sub-chapter of literature review.

Determining Best Model

By comparing the NRMSEs of different models the most suitable one is selected.

4 Chapter 4: Work Done so Far

This chapter will cover the progress done in first semester in accordance to objectives. There are two subchapter, (1) Classical Vs Neural Networks based landmark detection (2) Implementation of 2D Representative Approach. First sub-chapter will cover the theoretical differences between classical methods and Neural network (NN) methods to accomplish the first objective. Second sub-chapter will cover the implementation progress of 2D representative method based on cylindrical projection (Terada et al., 2018) and cascaded regression model for landmark identification algorithm (Kazemi & Sullivan, 2014).

4.1 Classical Vs Neural Networks based Landmark Detection

This sub-chapter will cover the overview of landmark detection techniques for classical and neural network approaches.

4.1.1 Classical Approaches

Most algorithms for facial landmark detection can be classified as either model-based or regression-based. Model-based approaches learn constraints on the arrangement of landmarks relatively to each other. Here the locations of landmarks are often constrained by other landmarks. The most common models are Active Shape Model (ASM), Active Appearance Model (AAM) and Constrained Local Model (CLM).

Regression-based approaches do not have an implicit shape model. They instead operate directly on the image and regress the coordinates of the landmarks. Cascaded regression approaches run multiple regression models consecutively, each one refining the predictions of its predecessor.

These are the most important classical approaches as was presented in (Taskiran et al., 2020) and (Johnston & Chazal, 2018).

4.1.1.1 Active Shape Models (ASMs)

Introduced in 1995 by (Cootes et al., 1995) , ASMs are an algorithm that place constraints on the locations of individual landmarks. Possible shape ,vector representation of an object, variations are learned from the training set and stored in a Point Distribution Model (PDM). PDM is a shape description technique that stores shape mean geometry and the geometric variations between shapes of same training set. Using Principal Component Analysis (PCA), variations in PDM are used to obtain a basis of shape parameters, from which new shapes can be generated.

When model is trained, Locating landmarks involves finding shape parameters that correspond to a shape as close as possible to the test image. The initial location estimates are set to be the mean of the training set. Then, iteratively a local region around each landmark is analyzed. This is then used to update the pose and shape parameters to best fit the new landmark locations. The shape constraint is maintained by only changing the landmark locations indirectly through shape parameters as opposed to directly changing them. This process is repeated until convergence.

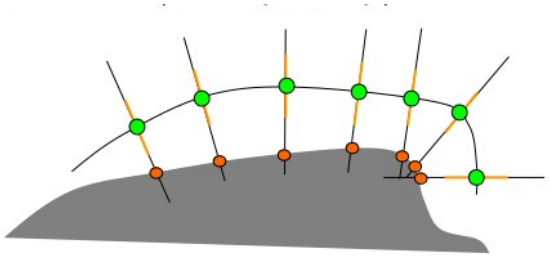


Figure 5: visualizing the ASM process. Red dots represent the test image. Green dots are estimate values of the shape of test image. The estimates are iterativly and indirectly changed to obtain best representation of test image. Best representation is one with lowest difference with the test image.

ASMs are fast simple,and can be efficiently extended to 3D to provide 3mm to 8mm accuracy (Smolyanskiy et al., 2014) . Furthermore there is a reasonable reliability based on data variation , 93.5% (Lee et al., 2015). However the method is not sufficiently accurate enough to be used to provide landmarks for registration, and any improvements may be unlikely as accuracy has remained at 3mm at best (Brehler et al., 2019).

4.1.1.2 Active Appearance Models (AAMs)

After success of ASM, the author introduced AAMs (Cootes et al., 1998). The main difference to ASMs is that texture variation around the landmarks is learned in addition to the shape variation. AAMs are appearance models, such models include the distribution of the data itself, and generates how likely a given example is. Landmarks in unseen images are located by finding shape and appearance parameters that minimize the difference between the generated image and the test image.

The shape model is built in the same way as in ASMs. The appearance model learns the appearance variations in shape-free images. A shape-free image is obtained by aligning the original image with the mean image. To generate a synthetic image, first the appearance parameters are used

to generate an image that is shape-free. Then the shape parameters are used to change it to the desired shape.

To find the optimal shape and appearance parameters, the error between the original image and the synthesized image is minimized. A key difference with ASM is that AAMs are a holistic model where they attempt to minimize difference with the whole image as opposed to an small local area.

ASMs are more accurate and faster than AAMs, but the latter are able to better match the texture. Furthermore AAMs suffer from lightning changes. For our project applications AAMs can be used on the 2D represented images. However this representation doesn't contain a lot of texture information, because the image is in gray-scale with little contrast. In another hand, 3D method of landmarking primarily contains shape information as opposed to texture information (refer to figure 9 to observe the lack of texture information). Another downside to AAMs is that they are holistic which would entail lower landmark accuracy.

4.1.1.3 Constrained local models (CLMs)

Proposed in (Cristinacce & Cootes, 2006) , CLMs are another shape models. CLMs are similar to AAMs as both model the shape and appearance variations of object. The main difference with AAMs is that, CLMs model the appearance around each landmark, whereas AAMs model the whole object. The appearances are modeled in local templates.

When using the CLM to locate landmarks, the shape and appearance models are used to generate estimated locations and texture templates around each landmark. Then for each landmark the correlation between the template and the actual image at this position is computed. The iterative fitting process maximizes these correlations while respecting shape constraints.

A member of CLMs family, the Convolutional Experts CLMs achieved a mean difference error of about 5.31mm for 68 landmarks (Zadeh et al., 2017) . Just like AAMs, the algorithm relies heavily on texture information and our project lacks such information.

4.1.1.4 Regression and cascaded regression

An alternative to shape based models is regression-based approaches. Regressors can be cascaded or cascaded.

An example of non-cascaded regression for facial landmark detection is use of random forest found in (Dantone et al., 2012) . There is one random forest for each of five head pose ranges and during inference the predictions of each random forest is weighted by the probability of this specific head pose. Each random forest is only trained on one specific head pose. This allows the random

forest to focus on the appearance specific for that head pose. In addition to the actual regressors for the facial landmarks, a regressor for the head pose probability is trained.

An example of cascaded regression in facial landmark detection is by (Shizhan Zhu et al., 2015). The cascaded regression patter, is used to search the shape spaces to obtain one that contains all shapes seen during training. In each stage the best fitting shape is picked and then refined in the next stage. This is done by restricting the search space to the most likely candidates. In the last stage the final predictions are computed. Their approach does not require an initial estimate for the location of the landmarks. Moreover, by starting from a coarse scale (considering all possible shapes with all possible poses) and refining the shape space on each cascade level, the algorithm is less prone to local minima. They use different features on each level. The first level makes use of less accurate features while the last stage uses more accurate features. Due computation quantity the first level is fast while second level is slow. Splitting the computation allows for real-time application.

Cascaded regression can provide better robustness against data variations compare to non-cascaded regression. Furthermore they can provide real time performances. Regressors can model complex appearance variance that is caused by different facial expressions, and face variability between individuals. This method is primarily being used on 2D data as it provides a fast means to normalization and noise reduction. It's application for MRI facial landmark detection is difficult to determine as despite providing less than optimum normalized RMS (0.49), the value can be further improved by averaging values of 3D coordinates from different face poses.

4.1.2 Neural Network Approaches

Classical facial landmark detection algorithms, such as the ones presented in the previous section, were implemented using manually engineered features and classical machine learning algorithms. Since the rise of artificial neural networks (NNs) most work in the field of facial landmark detection relies on Convolutional Neural Networks (CNNs).

In rest of this chapter NNs and CNNs are introduced. And then their feasibility for our project is determined.

4.1.2.1 Neural Networks

Neural networks are a powerful class of machine learning algorithms. For some applications like image processing, they do not require manually crafted features but are able to learn suitable features themselves. When using classic algorithms, features have to be engineered manually. This

could enable more accurate models since humans do not have to come up with optimal features for the task at hand. It also allows the use of complex hierarchical features because neural networks are typically organized in multiple layers that compute features on the features from the previous layer.

The main drivers for the success of neural networks are large amounts of available data and enough computational power to train these networks. Since these conditions were fulfilled, it has been possible to push the state of the art results by a large margin.

neural network can be seen as a complex non-linear function that maps the input (e.g. an image) to a set of output values (e.g. coordinates). Such a network usually consists of multiple layers, each containing a number of neurons. The neurons in each layer are connected to each neuron of the previous layer. There is a weight on each connection. The first layer is called the input layer and as such has no input connections. The last layer is the output layer and has no output connections. All layers between the input and output layer are hidden layers.

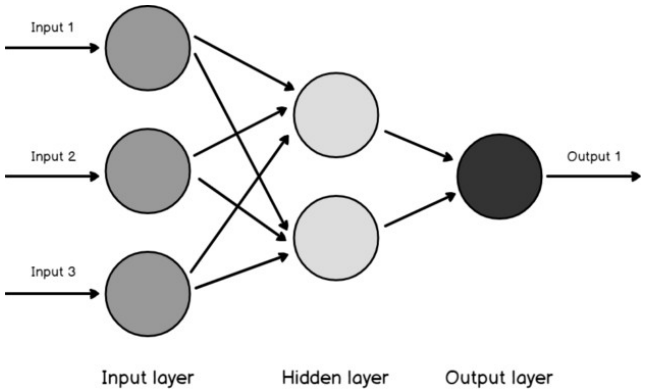


Figure 6: diagram of neural networks made up of neurons in input, output and hidden layers. number of neurons in input layer corresponds to number of inputs.

Networks such as Figure above, only allow data to be data to move forward, this kind of network is called feed-forward network. During training the weights of neurons are altered to minimize a loss function. This process is called back propagation. After training, the values of each neuron weights determines the relationship between input and output layers. Such values are representation of object feature.

4.1.2.2 Convolutional Neural Networks

In standard feed-forward neural networks all layers are fully-connected, which means that each neuron is connected to each neuron from the previous layer. However, this is not ideal when the input is an image, because images have a width, height and color channels. Furthermore, when analyzing images, a desired property is to have spatial invariance. This means that an object should

be recognized regardless of its position in an image. When using fully-connected layers this is not possible because there is a connection weight for each of the possible pixels.

CNN solve these issues by organizing the layers in three dimensions (width, height, channels) and replacing the fully-connected layers by convolutional, partial connected, layers.

4.2 Implementation of 2D Representative Approach

In the first semester a practical portion of FYP was carried out. One of the 2D methods of facial landmark detection, namely cascaded regression model (Kazemi & Sullivan, 2014), was used to locate facial landmarks in 3D. Before going through the model implementation, the data model of 3DSlcier, software of choice for medical image analysis, will be discussed.

Slicer Data Model

The Slicer Data Model is based on the Slicer Scene Data Structure. A Slicer scene is a collection of images, annotations, 3D models, spacial transforms, fiducials and cameras. Each element a scene is called a MRML node. The Medical Reality Markup Language (MRML) is an XML-based language used to serialize the content of Slicer scene on disk (scene.mrml). MRML is a data model developed to represent all data sets that may be used in medical software applications.

- **MRML software library:** An open-source software library implements MRML data in-memory representation, reading/writing files, visualization, processing framework, and GUI widgets for viewing and editing. The library is based on VTK toolkit, uses ITK for reading/ writing some file format, and has a few additional optional dependencies, such as Qt for GUI widgets.
- **MRML file:** When an MRML data is saved to file then an XML document is created (with .mrml file extension), which contains an index of all data sets and it may refer to other data files for bulk data storage. A variant of this file format is the MRML bundle file, which contains the .mrml file and all referenced data files in a single zip file (with .mrb extension).

MRML Scene

1 All data is stored in a MRML scene, which contains a list of MRML nodes.

2 Each MRML node has a unique ID in the scene, has a name, custom attributes (key:value pairs), and a number of additional properties to store information specific to its data type. Node types include image volume, surface mesh, point set, transformation, etc.

3 Nodes can keep references (links) to each other. For example a model node of skin can be linked to the volume node of head.

Basic MRML Nodes

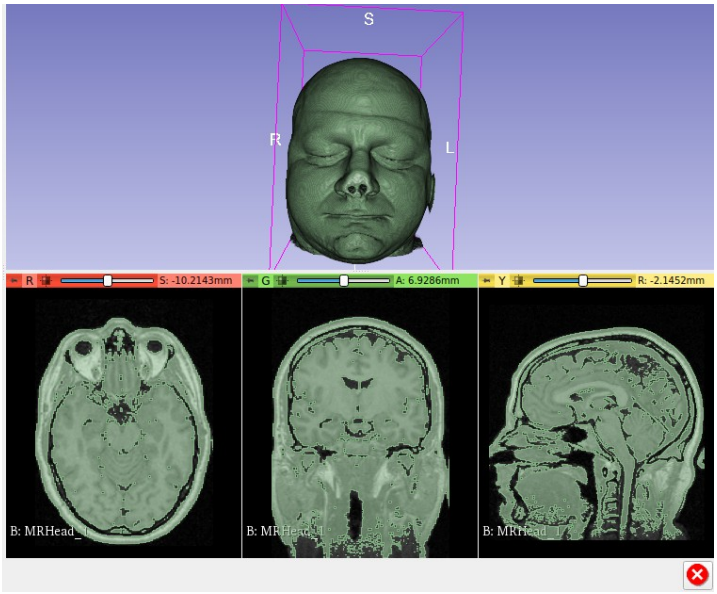
all objects in scene are stored in a hierarchical structure of MRML nodes. The basic types of nodes are as follows: (There are total of seven basic nodes)

1 Data nodes: store basic properties of a data set. Data nodes are typically thin wrappers over VTK objects, such as vtkPolyData, vtkImageData, vtkTable. Volume and Model are two examples of data nodes.

2 Display nodes: (vtkMRMLDisplayNode and its subclasses) specify properties how to display data nodes. For example, a model node's color is stored in a display node associated with a model node.

3 Storage Node: Describes how the data should be stored as file on disk. It can store one or more file name, compression options, coordinate system information, etc.

As the methodology workflow demonstrates the face had to be segmented from the head MRI. Using Segment Editor module of 3DSlicer ,accessed via python interactor, thresholding with intensity range of 40.70 to 279 was used to isolate the head of patient from the surrounding air. This resulted in a segment node that is linked to the volume node of head.



In order to separate the skin from the head a pipeline called the Wrap Solidify Effect is used. This pipeline is made up following functions:

1. A surface representation of voxels of the selected segment is created
2. Around this segment model, a sphere is created
3. To define the surface, iteratively the sphere model is shrieked and remeshed to match the surface of the segment model.

The result is a point cloud of the entire head.

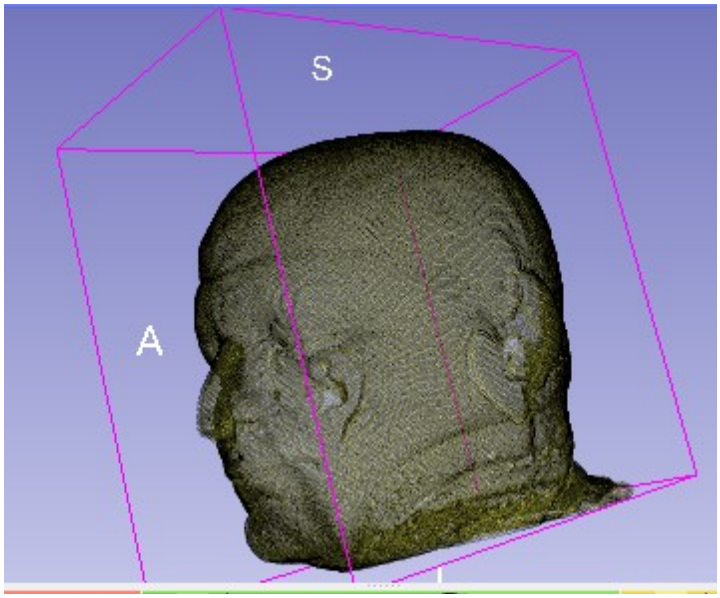


Figure 7: Point cloud of head skin

Next step is to clip only the face from this model. Given that we know most MRI images provide the head MRI centered at a particular location, we carryout clipping by simply accepting points that are located in a given rectangular box. Such boxes are represented by the 3DSlcier's cartesian coordinate system. The coordinate system has 3 axis, each with 2 different symbols.

In normal Cartesian coordinate system values along an axis can be either positive or negative, but for 3DSlicer axis have their own unique symbols. First axis has symbols R (right) and L (left), second axis has A and P, and the third axis has S and I symbols.

This box or ROI (region of interest) is made up of below planes:

1. Face plane: located on the axial plan with
2. back plane: opposite of face plance
3. Right Plane: plane on right side of head
4. Left plane: plane on left side of head

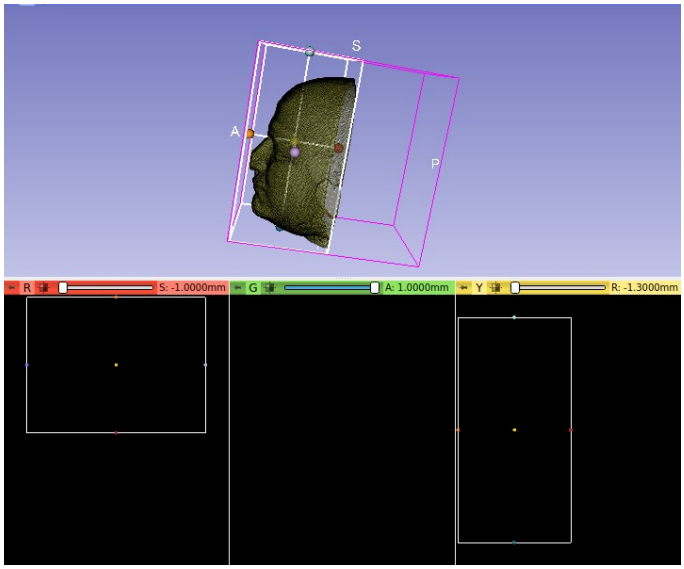


Figure 8: Clipping the face from head using ROI planes. The red box on bottom left represents the ROI plane located at Axial plane, while yellow box on bottom right shows the ROI plane on sagital plane.

Using the point cloud a cylindrical projection was made onto a cylinder of radius of 130 units. This could be done explicitly using the python interactor. Furthermore we can also obtain the ray casted image

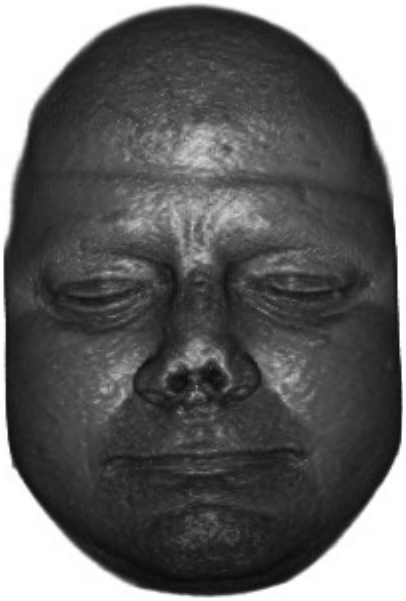


Figure 10: Orthogonal projection of ray traced surface

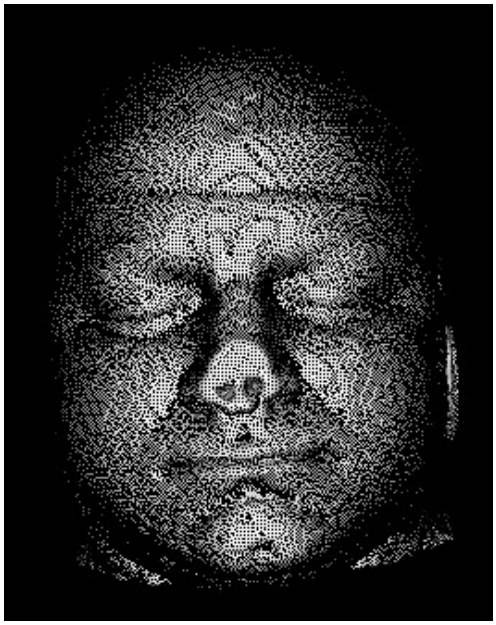


Figure 9: Orthogonal projection of point cloud where the distance away from the view plane is related to the intensity shown

We now feed these images into the pretrained cascaded regression model accessed using the dlib library. The results are as given:

Figure 11: Facial landmarks done by dlib on the depth map of the face

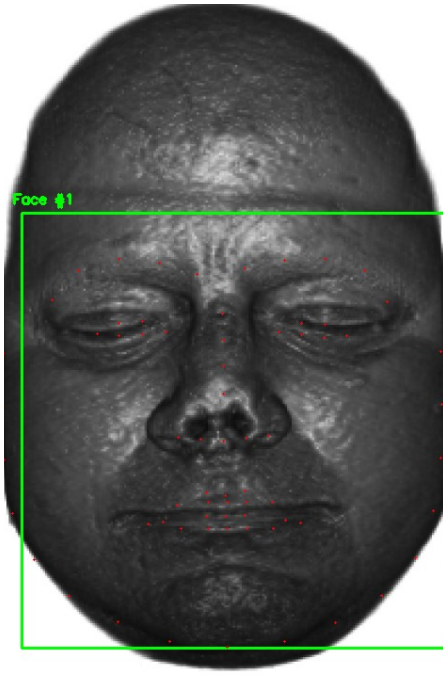


Figure 12: Facial landmarks obtained for ray casted image

Next step would be to convert the facial landmarks onto the 3D model. This will be carried out in the next semester

Discussion: The white dots on the depth map are actually points where there was no intensity values. This can be improved by increasing the sampling frequency or add additional points between the gaps. As far as accuracy is concerned the ray casted image provides better accuracy.

4.3 Planning for Next Semester

Process	Break	W1	**W2**	**W3**	**W4**	**W5**	**W6**	**W7**	**W8**	**W9**	**W10**	**W11**	**W12**	**W13**	**W14**
Complete and Optimize 2D methods															
Research: Classical Feature extractors															
Develop a NN 3D landmark detector															
Optimize the model															
Report writing															
Report submission															
Presentation															

5 Chapter 5: References

Brehler, M., Islam, A., Vogelsang, L., Yang, D., Sehnert, W., Shakoor, D., M.d, S. D., Siewerdsen, J. H., & Zbijewski, W. (2019). Coupled active shape models for automated segmentation and landmark localization in high-resolution CT of the foot and ankle. *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*, *10953*, 109530P. https://doi.org/10.1117/12.2515022

Çeliktutan, O., Ulukaya, S., & Sankur, B. (2013). A comparative study of face landmarking techniques. *EURASIP Journal on Image and Video Processing*, *2013*(1), 13. https://doi.org/10.1186/1687-5281-2013-13

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (1998). Active appearance models. In H. Burkhardt & B. Neumann (Eds.), *Computer Vision—ECCV'98* (pp. 484–498). Springer. https://doi.org/10.1007/BFb0054760

Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding*, *61*(1), 38–59. https://doi.org/10.1006/cviu.1995.1004

Cristinacce, D., & Cootes, T. (2006). *Feature Detection and Tracking with Constrained Local Models*. *41*, 929–938. https://doi.org/10.5244/C.20.95

Dantone, M., Gall, J., Fanelli, G., & Gool, L. V. (2012). Real-time facial feature detection using conditional regression forests. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2578–2585. https://doi.org/10.1109/CVPR.2012.6247976

Johnston, B., & Chazal, P. de. (2018). A review of image-based automatic facial landmark identification techniques. *EURASIP Journal on Image and Video Processing*, *2018*(1), 86. https://doi.org/10.1186/s13640-018-0324-4

Kazemi, V., & Sullivan, J. (2014). *One Millisecond Face Alignment with an Ensemble of Regression Trees*. 1867–1874. https://openaccess.thecvf.com/content_cvpr_2014/html/Kazemi_One_Millisecond_Face_2014_CVPR_paper.html

Lee, Y.-H., Kim, C. G., Kim, Y., & Whangbo, T. K. (2015). Facial landmarks detection using improved active shape model on android platform. *Multimedia Tools and Applications*, *74*(20), 8821–8830. https://doi.org/10.1007/s11042-013-1565-y

Liu, J. C., Zhang, L., Chen, X., & Niu, J. W. (2017). Facial landmark automatic identification from three dimensional (3D) data by using Hidden Markov Model (HMM). *International Journal of Industrial Ergonomics*, *57*, 10–22. https://doi.org/10.1016/j.ergon.2016.11.001

Lüsebrink, F., Sciarra, A., Mattern, H., Yakupov, R., & Speck, O. (2017). T1-weighted in vivo human whole brain MRI dataset with an ultrahigh isotropic resolution of 250 µm. *Scientific Data*, *4*. https://doi.org/10.1038/sdata.2017.32

Sheeba, A., & Manikandan, S. (2014). Image segmentation using bi-level thresholding. *2014 International Conference on Electronics and Communication Systems (ICECS)*, 1–5. https://doi.org/10.1109/ECS.2014.6892783

Shizhan Zhu, Cheng Li, Loy, C. C., & Tang, X. (2015). Face alignment by coarse-to-fine shape searching. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4998–5006. https://doi.org/10.1109/CVPR.2015.7299134

Smolyanskiy, N., Huitema, C., Liang, L., & Anderson, S. E. (2014). Real-time 3D face tracking based on active appearance model constrained by depth data. *Image and Vision Computing*, *32*(11), 860–869. https://doi.org/10.1016/j.imavis.2014.08.005

Taskiran, M., Kahraman, N., & Erdem, C. E. (2020). Face recognition: Past, present and future (a review). *Digital Signal Processing*, *106*, 102809. https://doi.org/10.1016/j.dsp.2020.102809

Terada, T., Chen, Y., & Kimura, R. (2018). 3D Facial Landmark Detection Using Deep Convolutional Neural Networks. *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 390–393. https://doi.org/10.1109/FSKD.2018.8687254

Wang, L., Alpert, K. I., Calhoun, V. D., Cobia, D. J., Keator, D. B., King, M. D., Kogan, A., Landis, D., Tallis, M., Turner, M. D., Potkin, S. G., Turner, J. A., & Ambite, J. L. (2016). SchizConnect: Mediating Neuroimaging Databases on Schizophrenia and Related Disorders for Large-Scale Integration. *NeuroImage*, *124*(0 0), 1155–1167. https://doi.org/10.1016/j.neuroimage.2015.06.065

Zadeh, A., Lim, Y. C., Baltrušaitis, T., & Morency, L. (2017). Convolutional Experts Constrained Local Model for 3D Facial Landmark Detection. *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2519–2528. https://doi.org/10.1109/ICCVW.2017.296