Dimensionality Reduction Techniques in Medical Data Analysis

High Dimensional Mathematics

Joan Lasenby

Signal Processing Group,
Engineering Department,
Cambridge, UK
and
Trinity College
Cambridge

jl221@cam.ac.uk, www-sigproc.eng.cam.ac.uk/ ∼ jl

25 May 2017

Overview

- Medical data: huge quantities of 1D, 2D or 3D (or even 4D) data.
- Often our aim is to diagnose abnormalities or to predict changes, as signs of deterioration or improvement.
- Presented with large amounts of data (potentially collected at 100Hz or higher!) we need to either extract features or to reduce the dimensionality of the data in an unsupervised fashion.

Overview

- Machine learning techniques have become extremely popular – inevitably they are being applied to medical data.
- One criticism that has been levelled at ML techniques is that they often fail to generalise and clinical practitioners have a very poor feeling as to why they work, thus causing scepticism.
- However, the new generation of ML systems are starting to be able to give the user clues as to exactly how they are learning.

Matrix/Tensor Decomposition

Dimensionality reduction via matrix/tensor decomposition can often provide features which give high performance classification, especially for time series data.

Singular Value Decomposition (SVD) is a factorization of a matrix X into the following form:

$$X = U\Sigma V^*$$

where U, V are unitary (orthonormal) matrices, V^* denotes the conjugate transpose (normally transpose for real data) of V, and Σ is a diagonal matrix containing the singular values of X – positive and in decreasing order.

SVD....

$$X = U\Sigma V^*$$

Note if $X : m \times n$, $U : m \times m$, $\Sigma : m \times n$ and $V : n \times n$.

Since we can write $XV = U\Sigma$, we see that X maps the columns of V onto the columns of U as follows:

$$X\mathbf{v}_j = s_j\mathbf{u}_j$$

Hence the analogy with eigen-decomposition.

The columns of U are often called the modes of X.

The columns of V are often called the loadings of X.

The magnitude of the singular values can be viewed as a measure of a mode's contribution to the matrix X.

SVD...

If we take $\hat{\Sigma}$ to be Σ with all singular values below a certain threshold replaced by zeroes [let there be p non-zero values], then a low rank approximation to X, which we call X_p, is given by

$$X_p = U\hat{\Sigma}V^T$$

In the above approximation, only the first p columns of U and V are used in the reconstruction.

Thus, since $U^T X = \Sigma V^T$, we can obtain a reduced feature set by multiplying our data by $U_{1:p}$, ie

$$X_{red} = U_{1:p}^T X \equiv \hat{\Sigma}V^T$$

where X_{red} has dimensions $p \times m$.

Example 1: Non-invasive Lung Function Monitoring

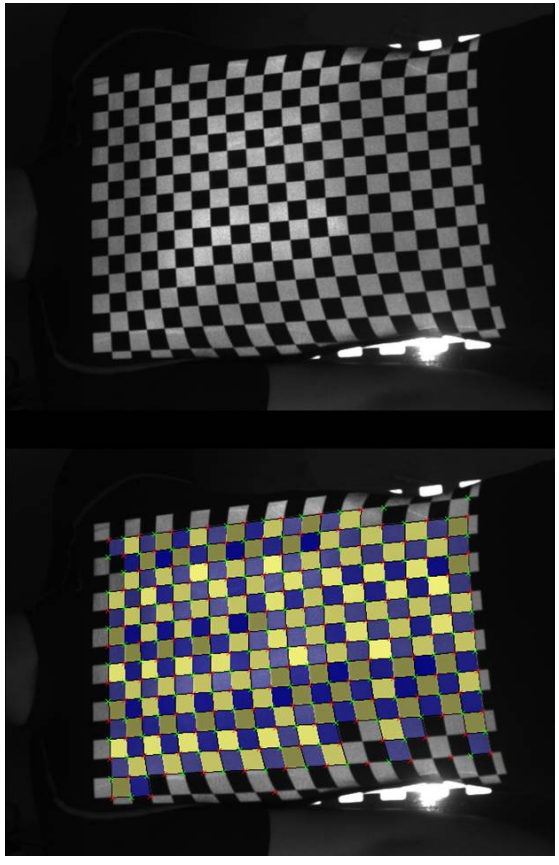
The most common form of Lung Function Testing is via Spirometry

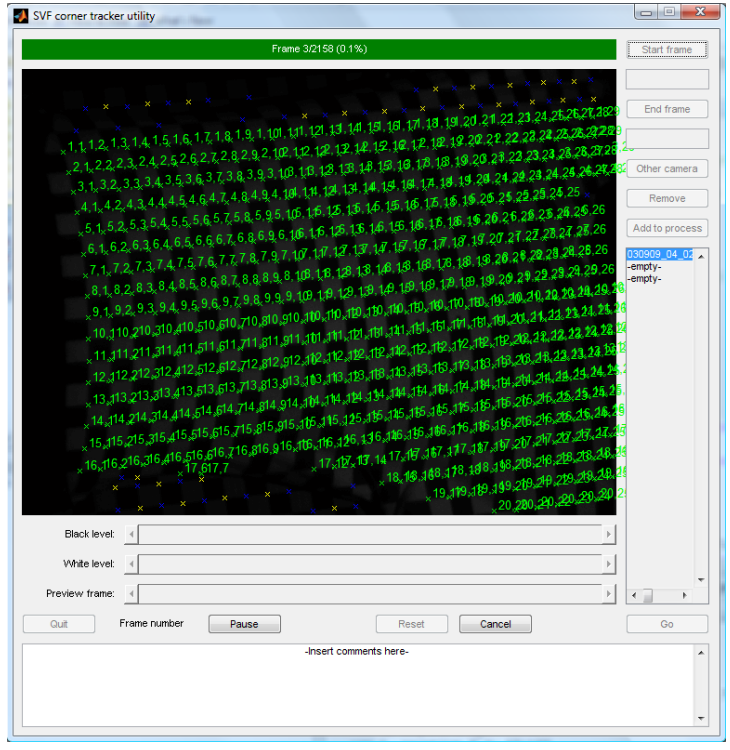
Spirometry is simple to use if the patient is able to perform the manoeuvres.

Unsuitability of Motion Capture for Infant Monitoring

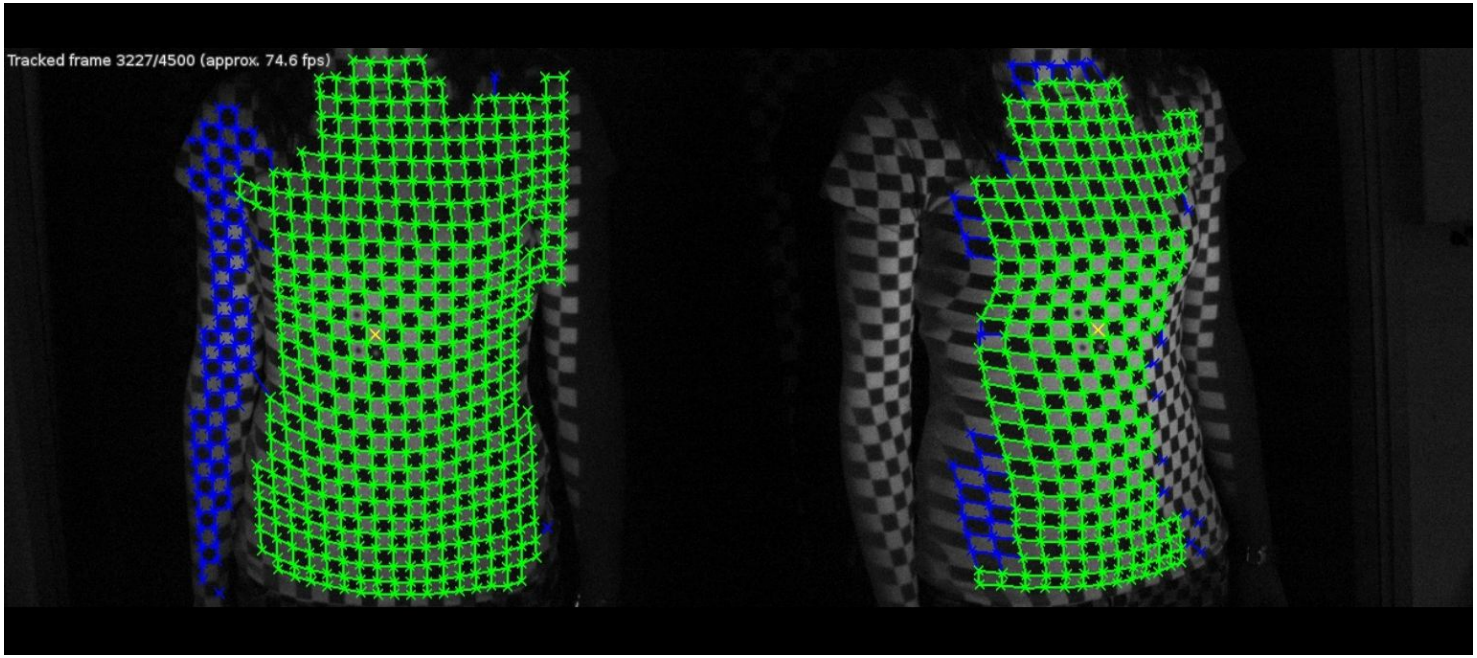
Development of a new structured light system: SLP Structured Light Plethysmography

Stage 1: tracking the grid in the two images

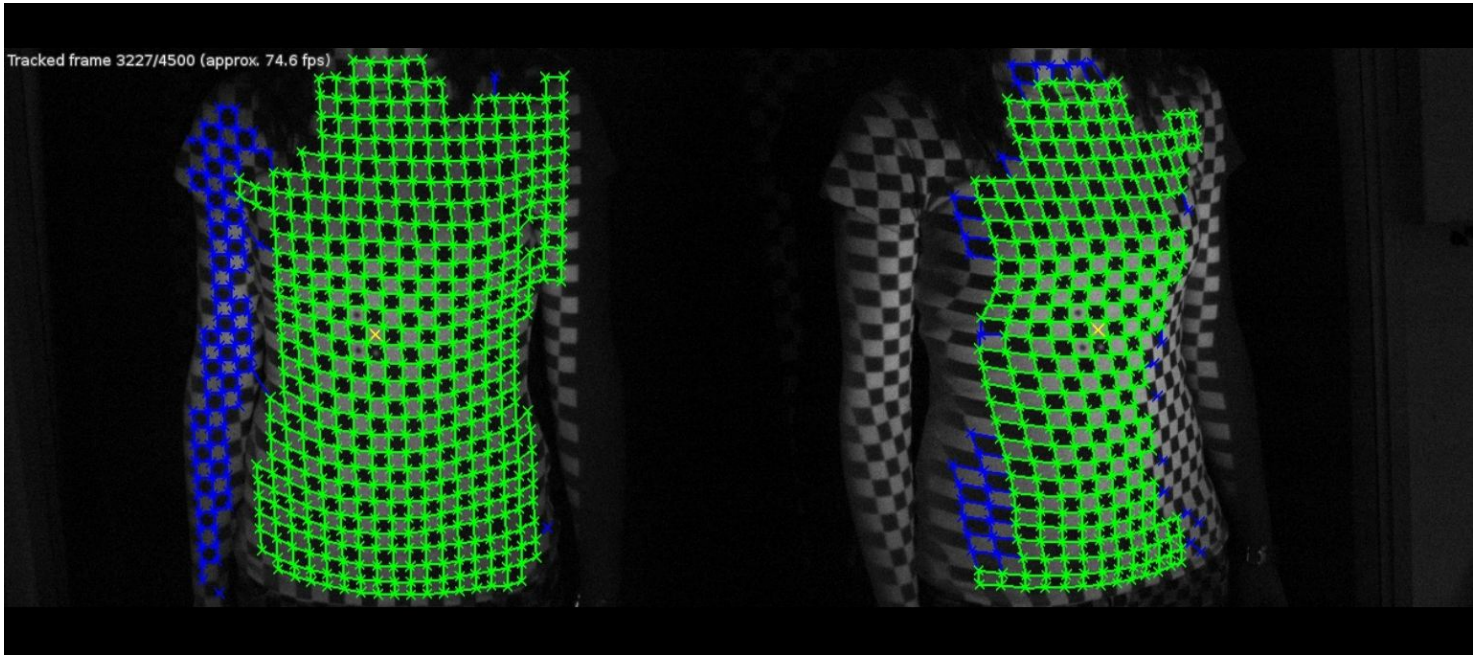




Stage 1: tracking the grid in the two images

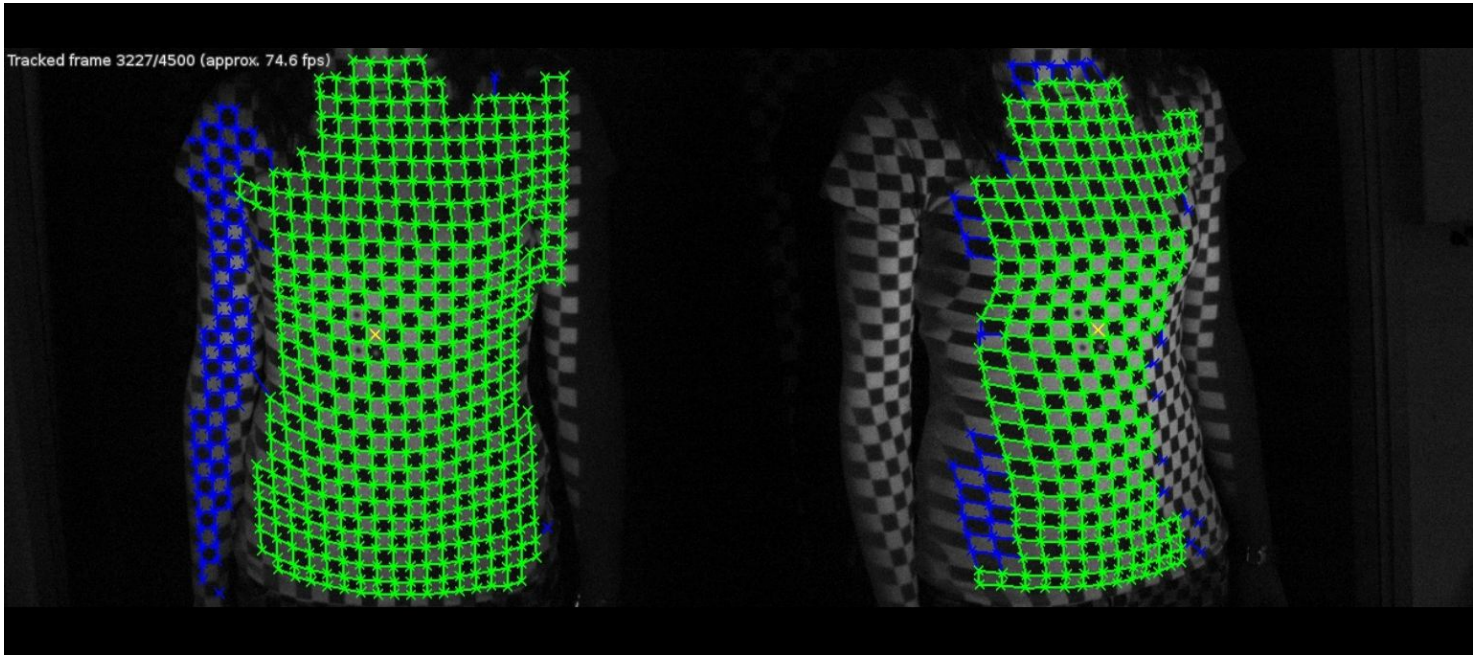


Stage 1: tracking the grid in the two images



Tracker is tailored to chessboard patterns and has been made robust to changes in lighting, texture (to some degree), orientation, distortion, creases, discontinuities etc.

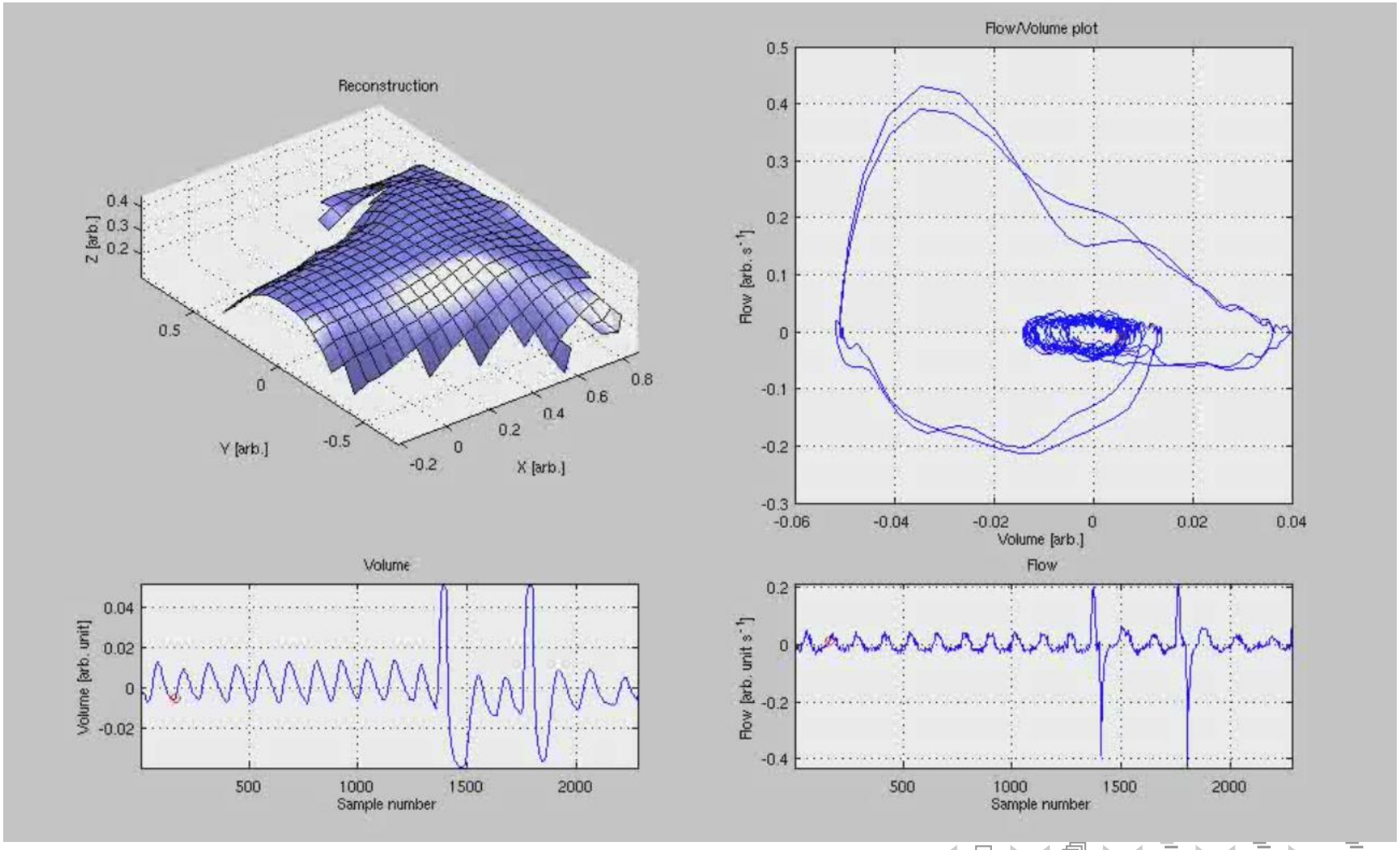
Stage 1: tracking the grid in the two images



Tracker is tailored to chessboard patterns and has been made robust to changes in lighting, texture (to some degree), orientation, distortion, creases, discontinuities etc.

It does the same thing as a Kinect but is more accurate!

Stage 2: dynamic chest surface reconstruction and volume calculation



Surface Decomposition of Lung Function Data

What our SLP method therefore gives us, is an accurate estimate of the chest/abdomen wall moving over time.

This can be for both forced manoeuvres and for tidal breathing.

Opens up the possibility of doing lung function evaluation on tidal breathing – using spatial as well as temporal information.

Therefore think about how we might start to characterise tidal breathing in terms of surfaces.

Regional Analysis

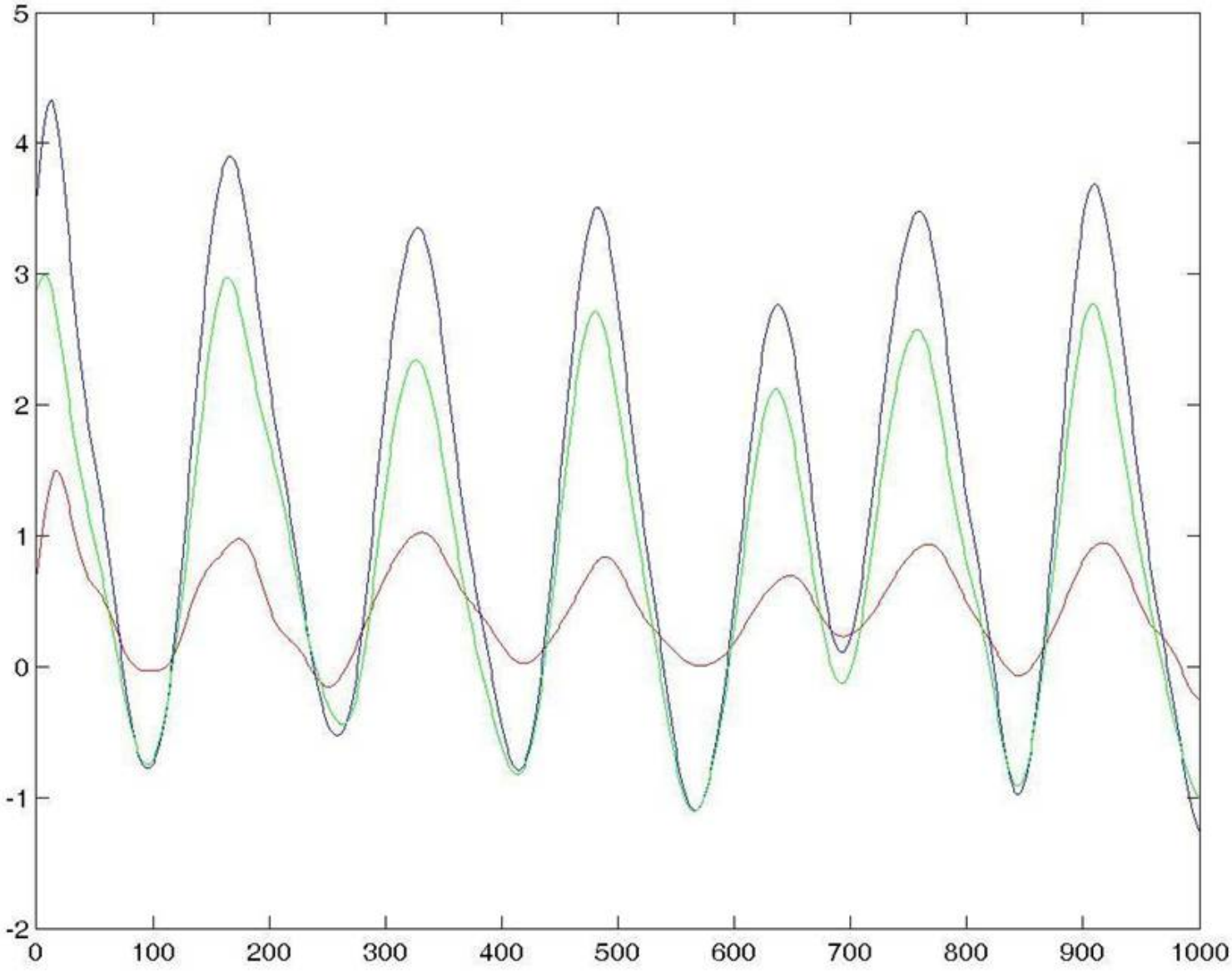
While we have verified that the SLP process is accurate relative to spirometry, we can now start to do things that spirometry cannot do.
One such thing is regional analysis

Regional Analysis

While we have verified that the SLP process is accurate relative to spirometry, we can now start to do things that spirometry cannot do.
One such thing is regional analysis

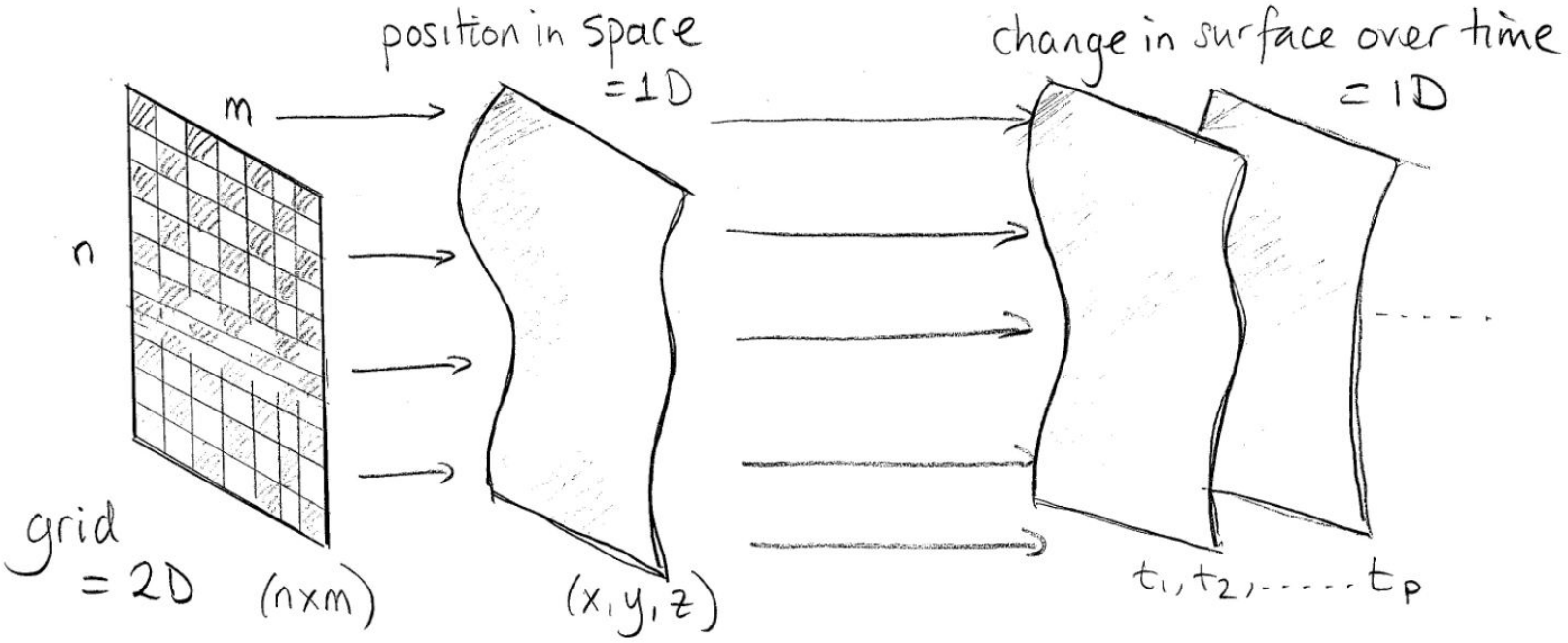
Regional Analysis

Fig 1: Tidal breathing perameter measured by SLP. Total respiratory Volume in blue, Chest volume in green and Abdominal volume in red.

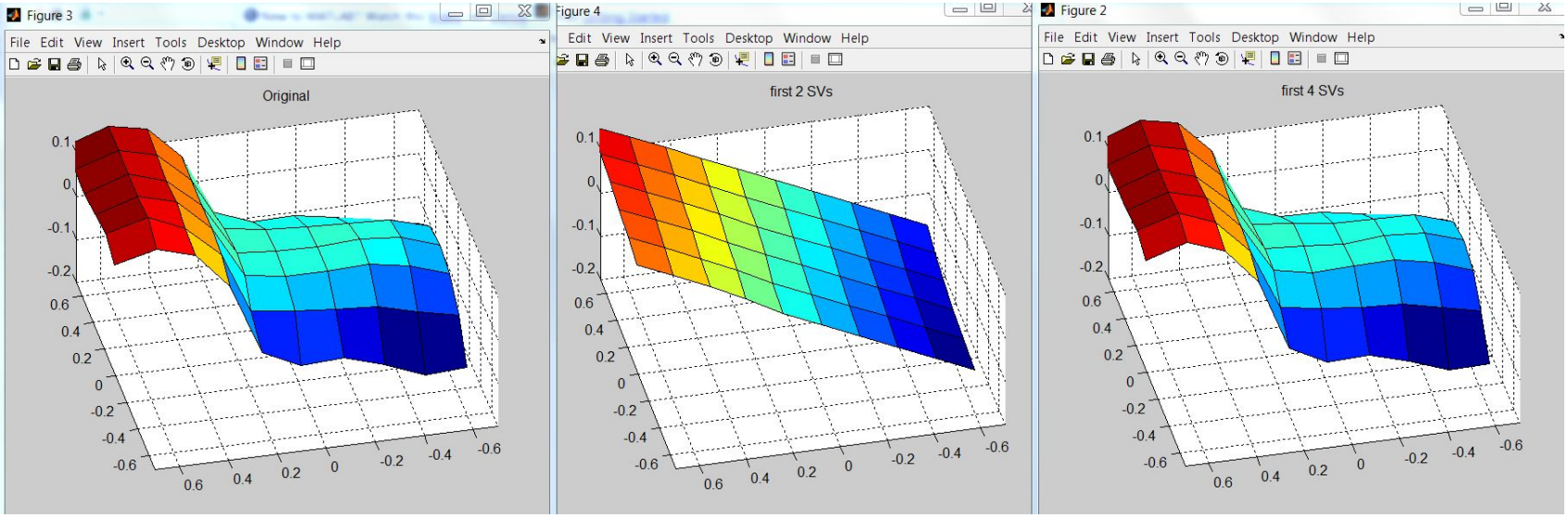


Surface analysis via Tensor Decomposition

In this breathing data case (a 4D dataset), the decomposition which works best is that of mapping 2D to 2D – ie mapping the grid-space onto the position-time space.



Surface Reconstruction



Characterisation of Disease

Normal tidal breathing seems to be characterised by:

- A large degree of variability in the time taken for individual breaths (can get this from SLP or spirometry)
- and
- A surface motion that can be characterised by few 'eigen-surfaces'

Asthma, COPD, etc seems to be characterised by:

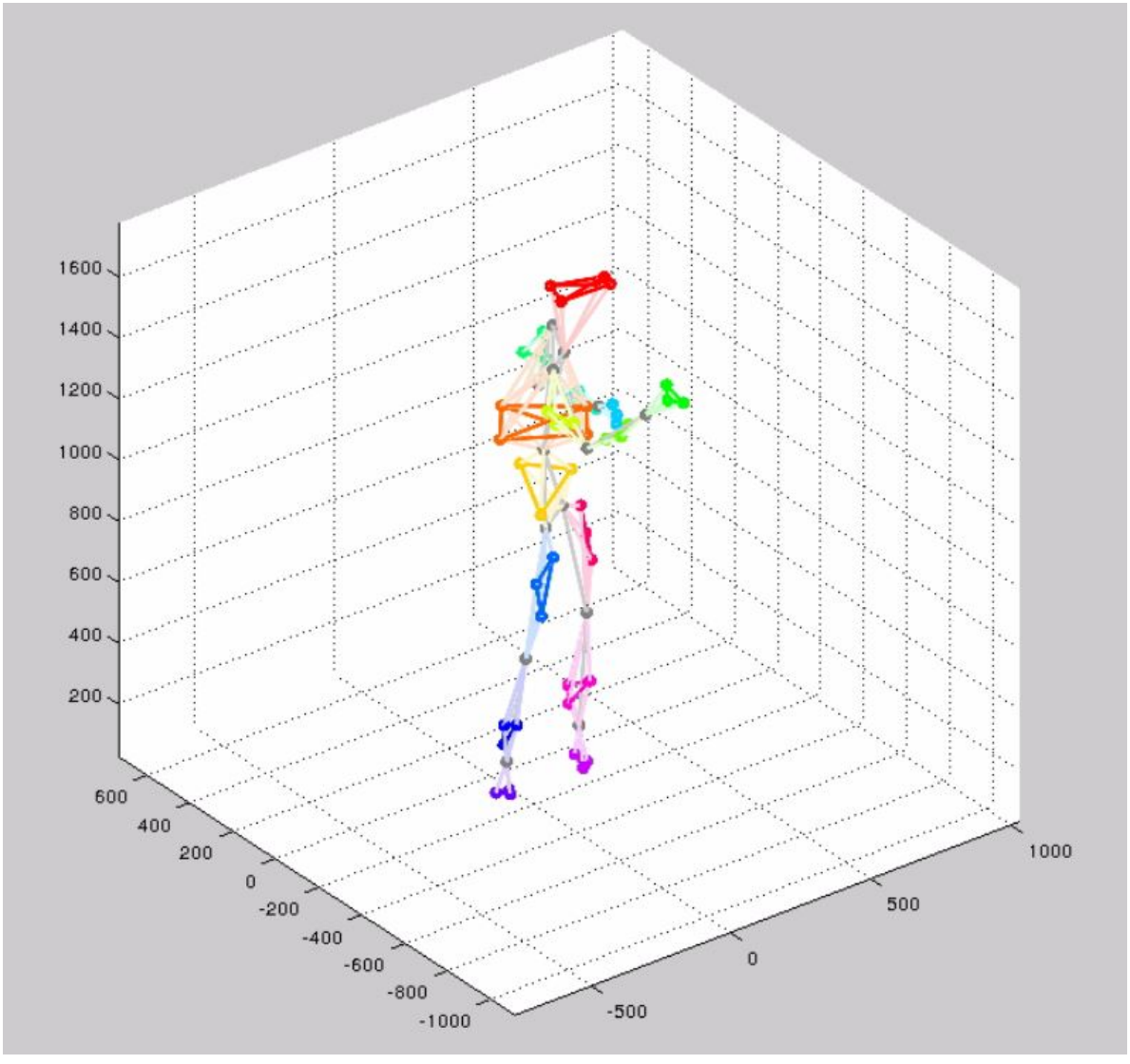
- Low variance of time taken for individual breaths.
- A surface motion that is more complex, needing more eigen-surfaces to describe it.

Example 2: Motion Analysis for Rehabilitation

Aim here is to:

- Automatically fit skeletons to optical motion capture data (involving multiple cameras and markers rather than Kinect-like systems) in order to save clinicians time (currently mostly they fit a predetermined marker set)
- Fit an accurate skeleton with a minimal marker set
- Break the data down into 'eigentrajectories' and define a data-driven approach to distinguishing between gait patterns (especially in stroke recovery).

Motion Capture Data



Eigentrajectories

The position of marker m in frame f is given by

$$\mathbf{r}_m^f = [x_m^f, y_m^f, z_m^f]^T$$

The trajectory of marker m is an $F \times 3$ vector [F is total number of frames]

$$\mathbf{t}_m = [\mathbf{r}_m^1, \mathbf{r}_m^2, ..., \mathbf{r}_m^F]^T$$

Now stack up the trajectories for each marker and perform SVD/PCA:

$$\mathbf{T} = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ \mathbf{t}_1 & \mathbf{t}_2 & \cdots & \mathbf{t}_M \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}$$

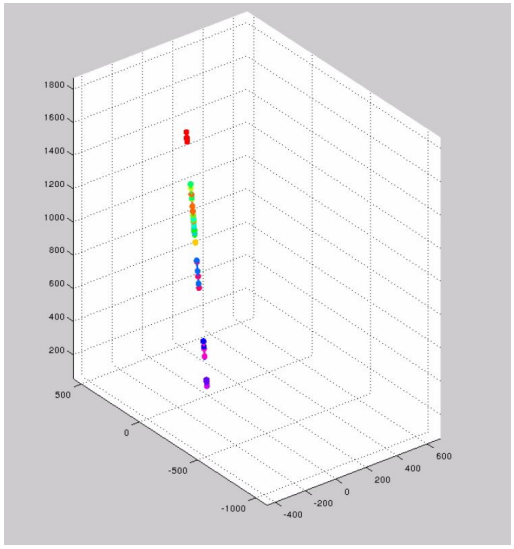
$$\mathbf{T} = \mathbf{U} \quad \boldsymbol{\Sigma} \quad \mathbf{V}^\top$$

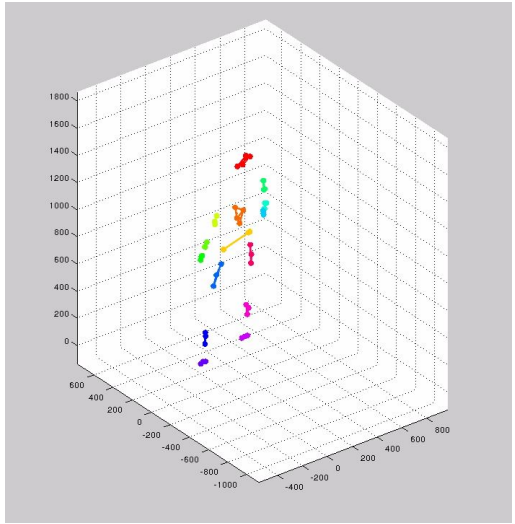
Eigentrajectories....

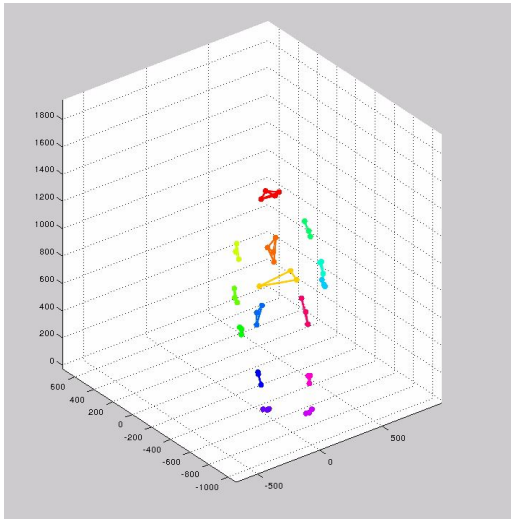
The columns of U are the eigentrajectories.

The eigentrajectory space is the column space of T – any of the trajectories can be built up as a linear combination of the eigentrajectories.

We can use these ideas to extract accurate estimates of the joints, and then look at the eigentrajectories of the joints to characterise gait.





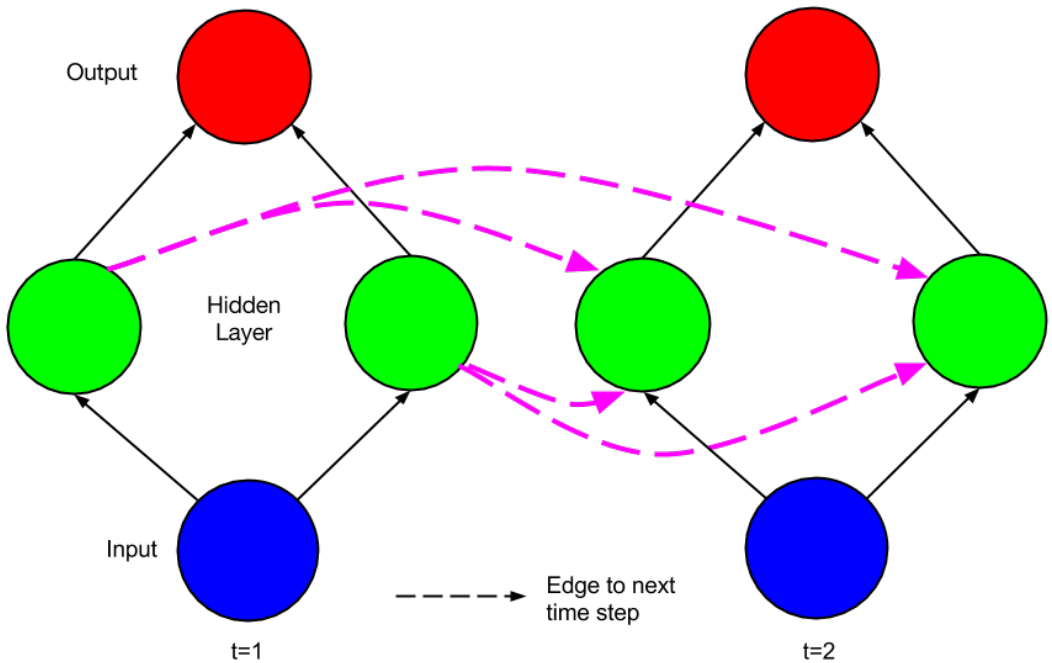


Machine Learning (ML) for Medical Datasets

- There are many very large 1D medical datasets (usually consisting of long time series for many subjects/patients) currently available: EEG, ECG, PSG, SpO2, etc...
- There are now also datasets which provide classification, so researchers can test out their ML success rates. eg PhysioNet: `https://www.physionet.org/`
- Here we will briefly just look at one form of ML –Recurrent Neural Networks, [RNNs]

Recurrent Neural Networks

RNNs are vanilla Neural Networks unfolded in the time domain as illustrated below. The RNNs have the hidden (green) units connected between contiguous time steps, and the input can be fed into the network in a per-time-step fashion.



[Adapted from Lipton, 2015]

Recurrent Neural Networks

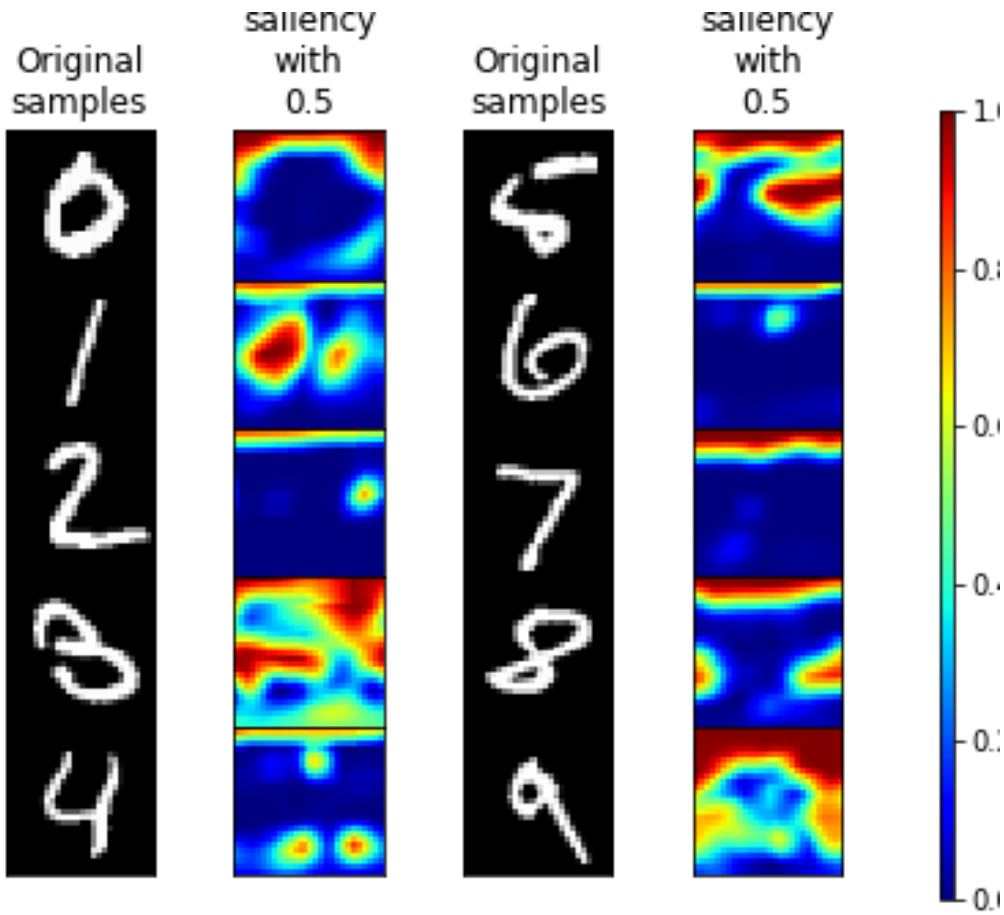
- Recurrent Neural Networks are renowned for being uninterpretable black boxes.
- In many domains, such as medicine, it is imperative to understand the decisions made by such models.
- We have been using visualization techniques to elucidate some mechanisms of RNNs. [J. van der Westhuizen and J. Lasenby, Visualizing LSTM decisions, 2017: `http://arxiv.org/abs/1705.08153`]

Visualisation

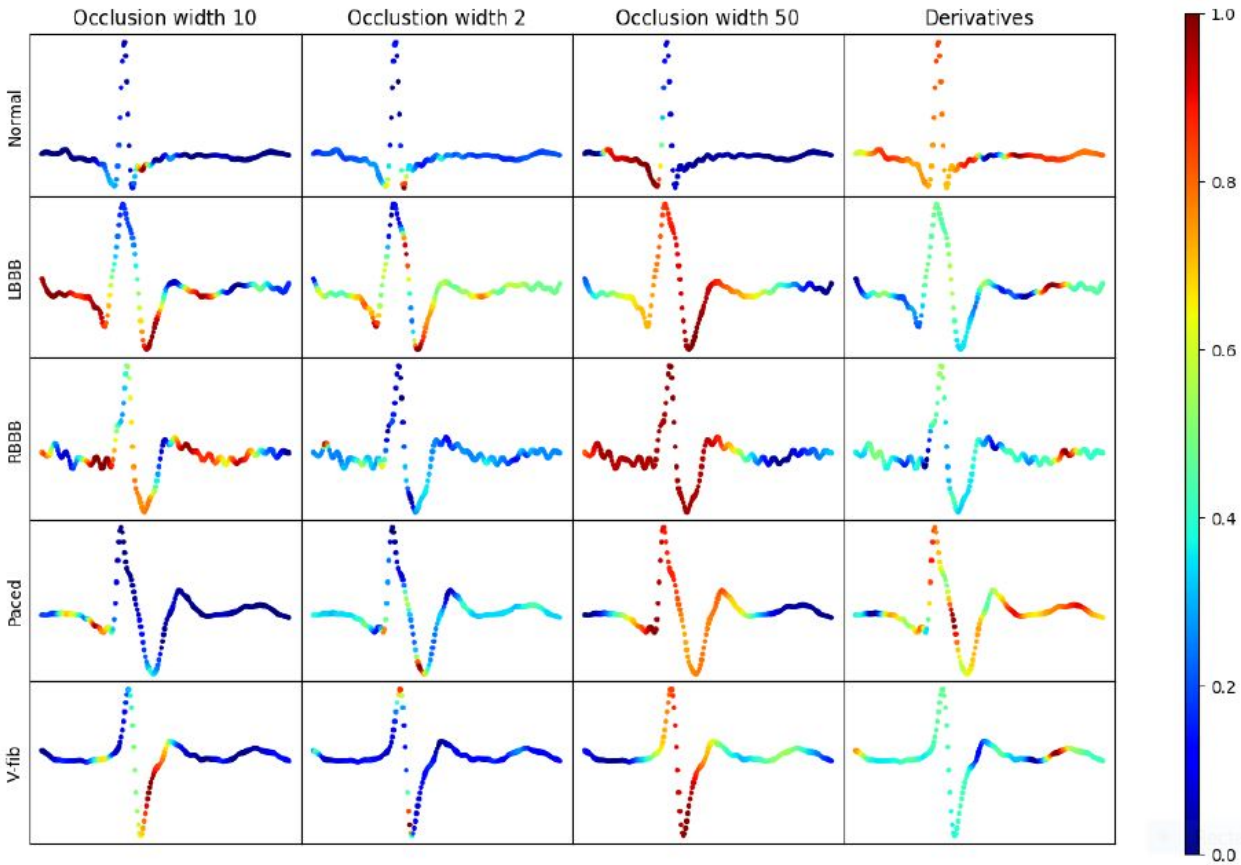
- Input feature saliency: To illustrate what input features are important we iteratively occlude sections of the input over time.
- The predictive scores are summed over the iterations and normalized to yield a heat map of the time steps that were the most important for classifying a signal into a specific class.
- We show examples of this occlusion technique on the MNIST handwritten digit dataset where occlusions were performed in 5x5 blocks

Visualisation cont...

In the image below, the importance of each input step is shown on a scale 0 to 1, with 1 being the most important. Original samples are shown next to the input feature saliency.

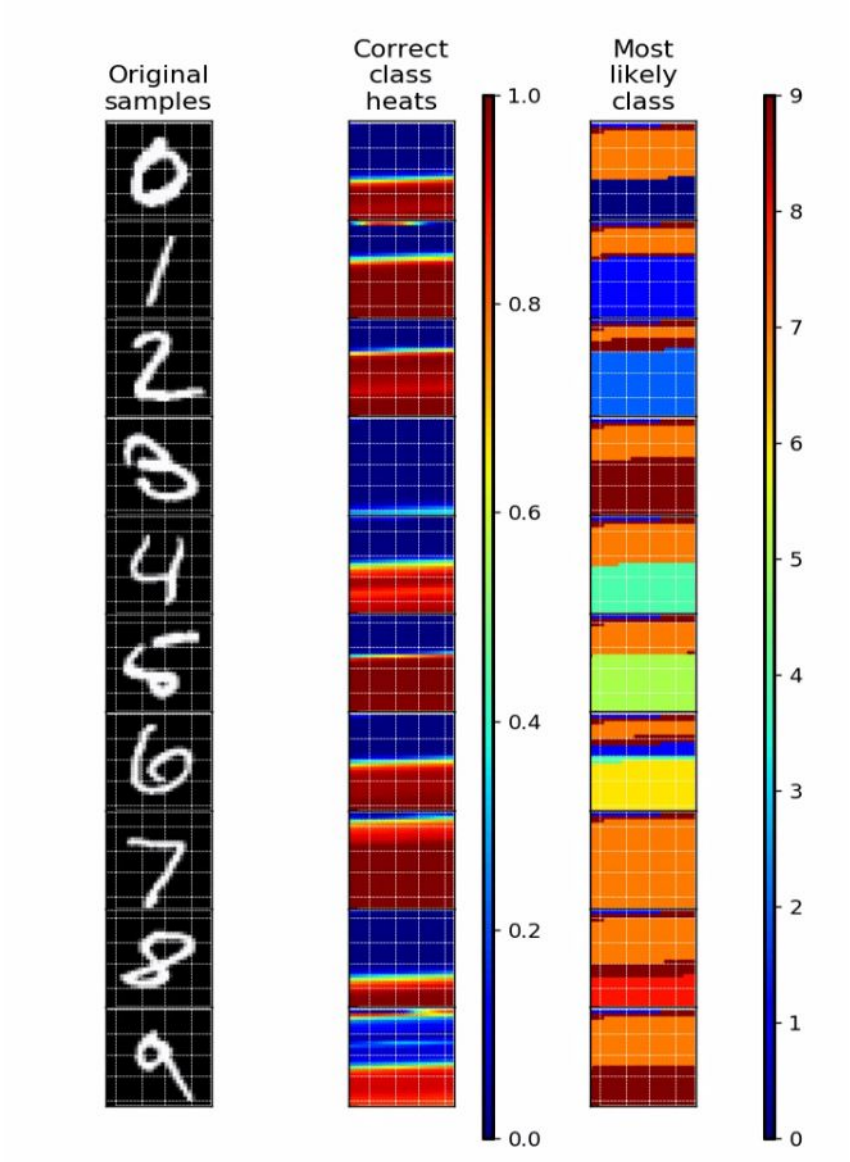


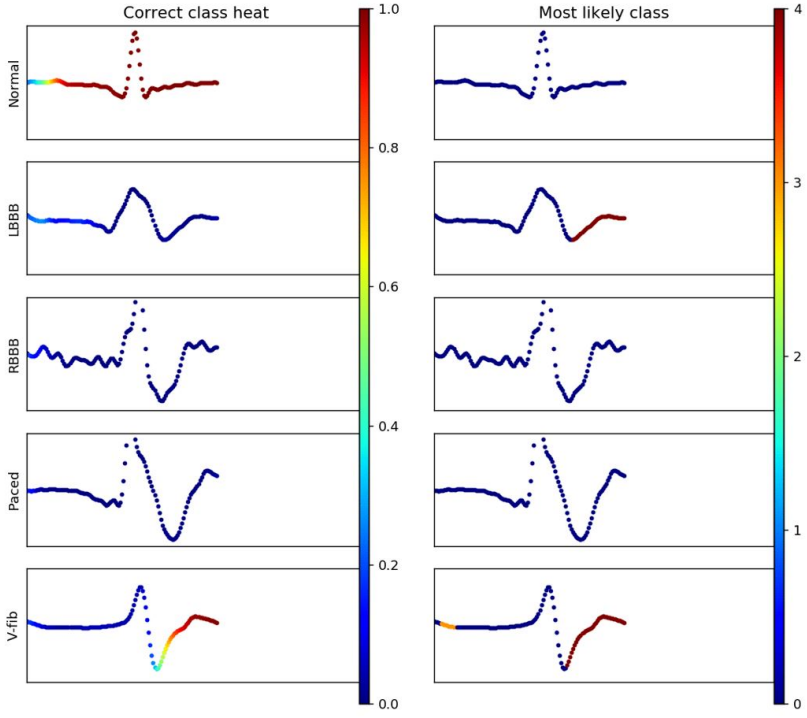
Applying this visualisation to medical time series



This figure shows the importance of each input step on a scale of 0 to 1.[All were classified correctly].

More visualisation....





Summary

- Often our data will need some sort of pre-processing before any form of classification (eg – make 'time series' the same length or normalise in other appropriate ways).
- Finding the modal structure of a given dataset usually gives us useful information and can complement specific feature extraction systems, or be a front end for such systems.
- There has been much recent progress in visualising what neural networks are doing as they learn.
- Currently the best performing networks (on most datasets) are those that use hand-crafted features as input. An understanding of what ML techniques are learning may improve generalisation.