Shape Features

Audiovisual Processing CMP-6026A

Dr. David Greenwood

david.greenwood@uea.ac.uk

SCI 2.16a University of East Anglia

November 5, 2021

Content

- Speech recognition models
- Visual Features
- Image segmentation
- Point distribution models
- Fourier descriptors

Acoustic Speech Recognition

The task of a speech recogniser is to determine the most likely word sequence given a new sequence of (acoustic) feature vectors.

Acoustic Speech Recognition

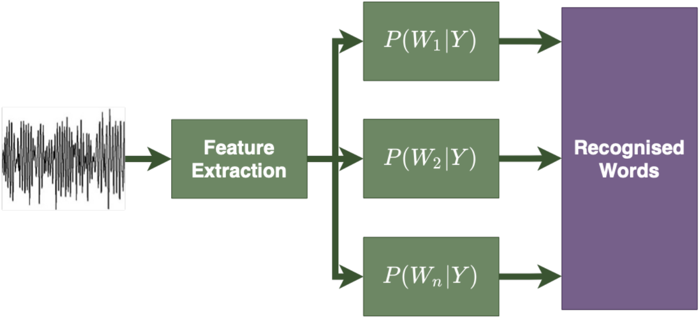
An elegant way to compute this is using hidden Markov models.

$$P(W|Y) = \frac{P(Y|W)P(W)}{P(Y)}$$

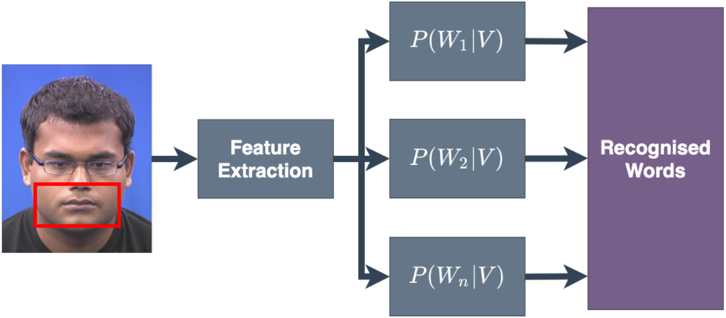
Acoustic Speech Recognition

Learn the parametric model from training data, and use to estimate the probabilities.

Acoustic Speech Recognition



Visual Speech Recognition



Audio-Visual Speech Recognition

Combine two modalities using:

- Late Integration
- Early Integration

Late Integration

Late integration builds two separate models and weights their probabilities to provide the recognised word sequence.

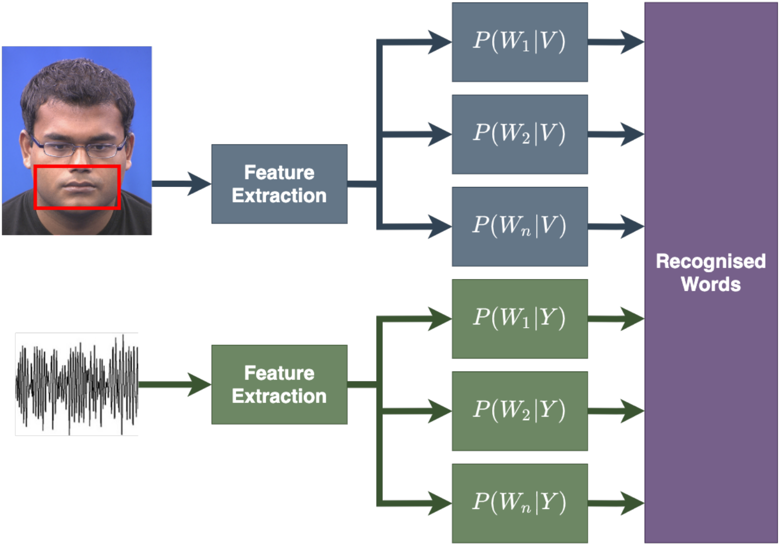
Late Integration

Has been shown to offer better performance than early integration.

Not straightforward to weight output probabilities.

An investigation of HMM classifier combination strategies for improved audio-visual speech recognition. Lucey et al. 2001

Late Integration

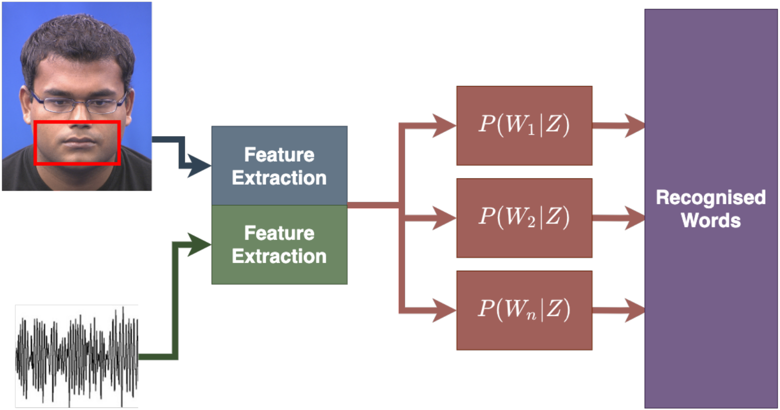


Early Integration

Concatenate the acoustic and visual models to form a single model.

Visual features often need **interpolation** to align with the acoustic features.

Early Integration



Visual Features

MFCCs are the standard features used in acoustic speech recognition.

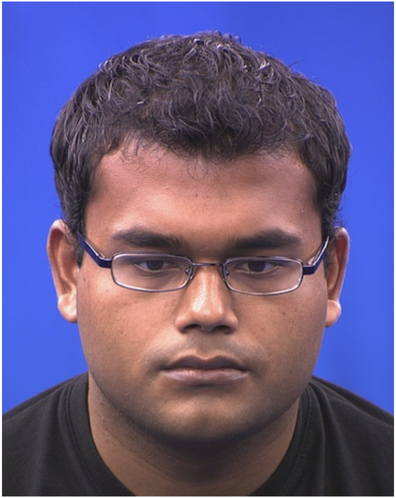
- What is the equivalent for visual speech?
- In short: there is little agreement!

Visual Features

Typical features include:

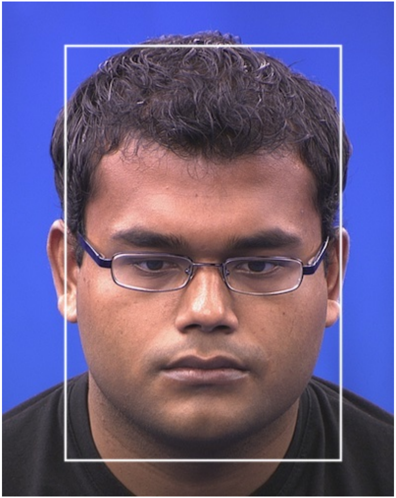
- Shape-based features
- Appearance-based features
- Hybrid features

Region of Interest (ROI)



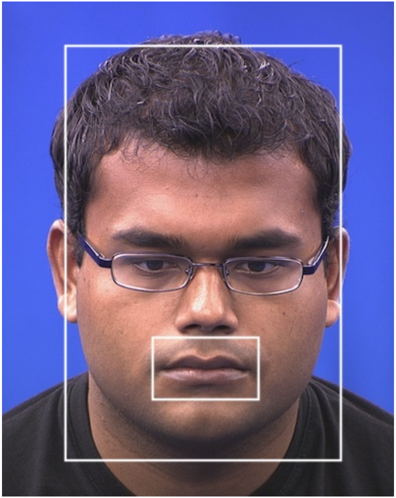
For any form of visual feature extraction, some form of localisation is required.

Region of Interest (ROI)



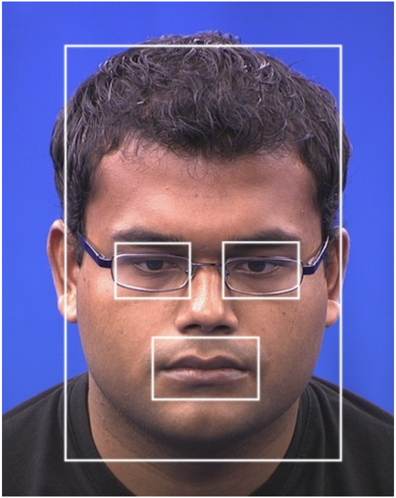
Where in the image is the face?

Region of Interest (ROI)



Where are the facial features of interest?

Region of Interest (ROI)



MATLAB has an implementation of the Viola Jones face tracker.

Shape Features for Recognition

Shape features *might* include:

- Articulatory-based features, such as mouth height and width
- Point distribution model (and related features)
- Fourier descriptors

Shape Features for Recognition

There is a trade-off between ease of extraction and the amount of information extracted.

- Sparse point sets are easier to locate, but capture less information
- Denser point sets are information rich, but require a more sophisticated capture process

Representing Shapes

We need a method for describing specific shapes in images.

Representing Shapes

An edge detector will locate edges in an image.

Representing Shapes

Which belong to the object of interest?
How are these allowed to vary as the object deforms?

Representing Shapes

Idea

Can we represent shapes using the image coordinates of the edge pixels?

Representing Shapes

We could, but the same shape in two locations will have different coordinates.

The coordinates describe the shape in the image coordinate frame, so they encode the shape **and the location** of the shape.

Representing Shapes

We are not interested in **where** the shape is — just the shape itself.

– A lip-reading system might use the shape of the lips to recognise speech, but it should not matter where in the image the lips are.

Representing Shapes

The primary problem is how to segment the lips from the background to extract a representation of the shape that is independent of image location.

A pre-processing stage of feature extraction identifies the region of the image that corresponds to the mouth.

This results in a binary mask, which is 1 if a pixel represents the mouth and 0 otherwise.

Image Segmentation

The goal of image segmentation is to classify each pixel as being either foreground or background.

Image Segmentation

We require three things:

1. A property that we can measure from the image pixels (e.g. colour).
2. A distance measure that defines how close two pixels are given that property.
3. A classifier that can discriminate one class from another using that distance.

Image Segmentation

Which colour-space should be used?

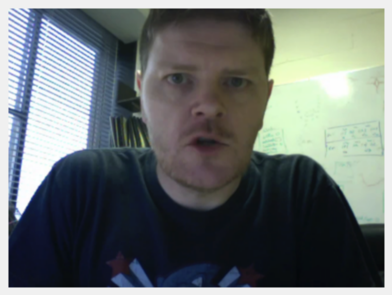


Figure 1: RGB

Image Segmentation

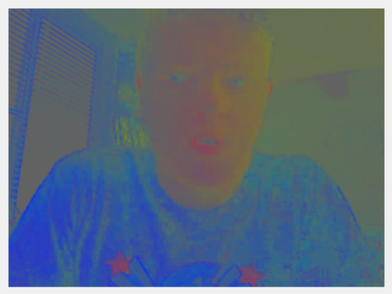


Figure 2: Normalised RGB

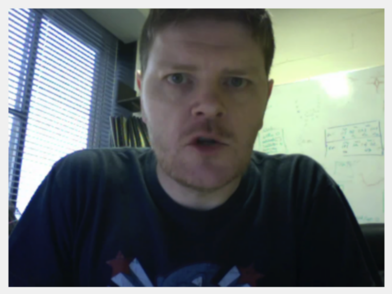


Figure 3: RGB

Image Segmentation

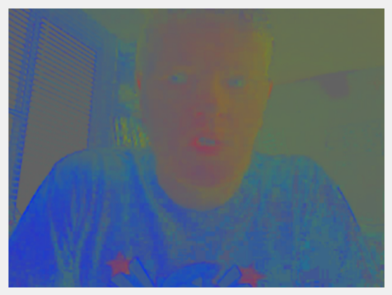


Figure 4: Normalised RGB

$$I = \left[\frac{r}{r+b+g} \frac{g}{r+b+g} \frac{b}{r+b+g} \right]$$

- A colour is represented by its proportion of red, green and blue, not the intensity of each.
- Reduces distortions caused by lights and shadows in an image.

Image Segmentation

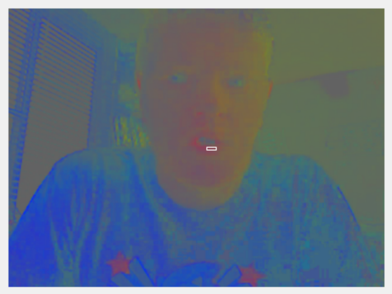


Figure 5: Normalised RGB

What colour do we want to segment out?

Image Segmentation

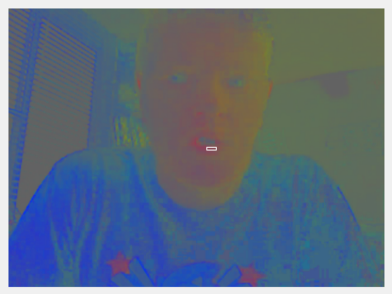


Figure 6: Normalised RGB

Find the mean colour of a lip pixel:

$$\begin{bmatrix} \mu_r \\ \mu_g \\ \mu_b \end{bmatrix} = \mu_c$$

Image Segmentation

Find the Euclidean distance between each pixel in the image, $I_{i,j}$, and the mean lip pixel colour μ_c.

$$D_{i,j} = \sqrt{\sum (I_{i,j} - \mu_c)^2}$$

Image Segmentation

$$D_{i,j} = \sqrt{\sum (I_{i,j} - \mu_c)^2}$$

A better distance metric might consider the variance of the lip pixels rather than just the mean, e.g. *Mahalanobis* distance.

Image Segmentation

Threshold the distance to segment lips from the background.

$$T_{i,j} = \begin{cases} 1 \; \mathit{if} \;\; D_{i,j} < \tau \\ 0 \; \mathit{otherwise} \end{cases}$$

Image Segmentation

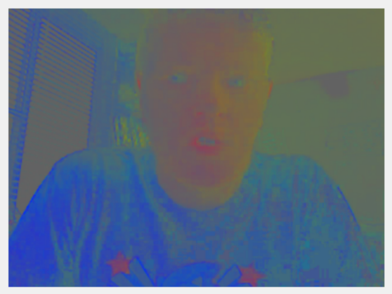


Figure 7: Normalised RGB

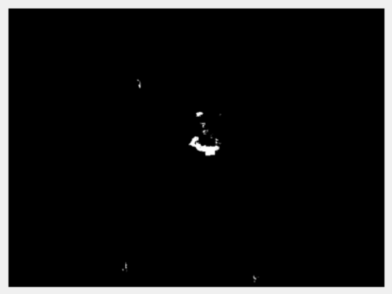


Figure 8: Threshold Image

Image Segmentation

- This approach assumes that there is nothing in the image that is the same colour as the lips, otherwise there is nothing to tell these regions apart.
- Often do other pre-processing (e.g. Viola-Jones face detector) first.

Image Segmentation

- Need to set the threshold, which itself is not trivial.
- If the threshold is too low, lip pixels will be missing.
- If the threshold is too high, background will be accepted as foreground.

Image Segmentation

The matte will still contain spurious pixels, which might need cleaning up using **morphological** filtering.

Image Segmentation

- From the resultant binary mask, the relevant features still need to be extracted.
  - e.g. articulatory-based features...

Articulatory-based Features

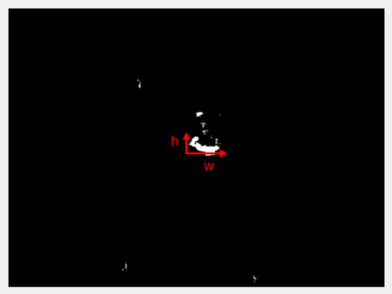


Figure 9: Threshold Image

From the binary image we can extract features such as:

- the height and width of the mouth region
- the number of pixels within the mouth region
- the mouth centroid

Image Segmentation

Automated approaches are attractive as there is no manual effort.
However:

- The colour of the lips is often similar to the surrounding skin.
- Noise is an issue.
- The facial appearance can change over time (e.g. beard growth, etc.).

Image Segmentation

Semi-automated approaches are generally more robust.

- They might need significant effort to reliably construct the model.
- But priors can be imposed on the expected shape.

Point Distribution Models

A *generative* statistical model of the variation of the shape of an object.

Point Distribution Models

Use **Principal Component Analysis (PCA)** to model the variation in the coordinates of a set of *landmark* points.

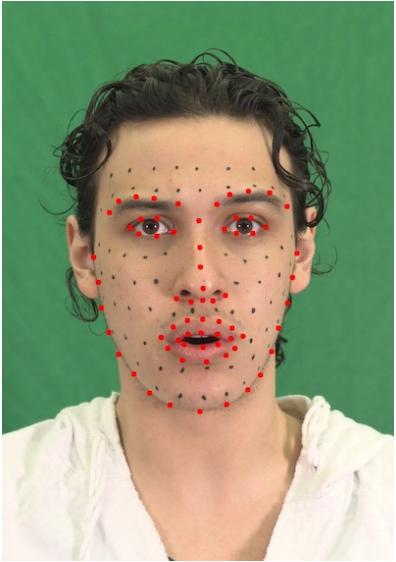
The PDM can represent complex shapes with just a few parameters.

Point Distribution Models

You can use an *Active Shape Model (ASM)* or *Active Appearance Model (AAM)* to automatically locate the landmarks (facial tracking).

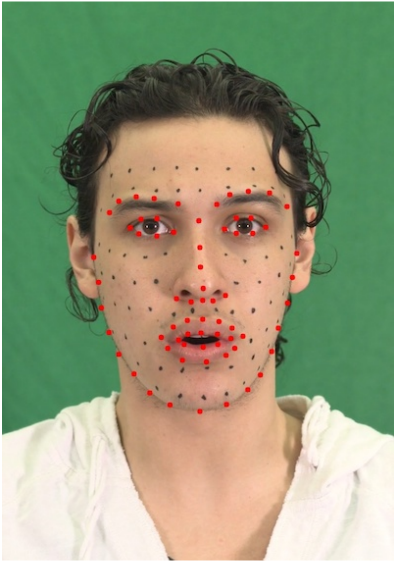
This model **requires training**.

Point Distribution Models



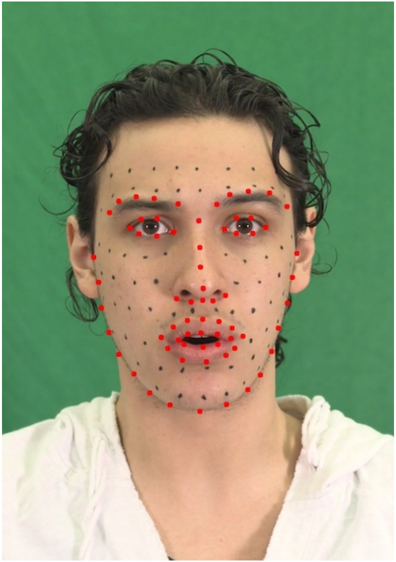
A **shape** is represented by a set of **landmarks** located along the shape boundary.

Point Distribution Models



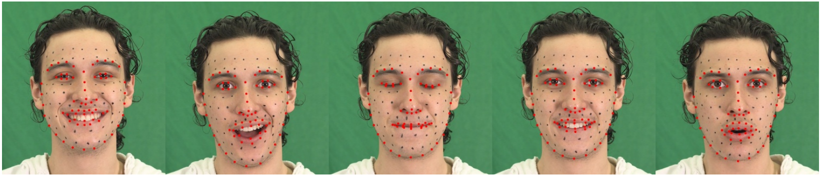
- The landmarks must be easy to locate from one image to another.
- T-junctions, points of high curvature, corners etc. form good candidates.
- Include evenly spaced intermediate points along the boundary.

Point Distribution Models



I provide a tool to annotate landmarks here: https://github.com/davegreenwood/face-landmark-tool

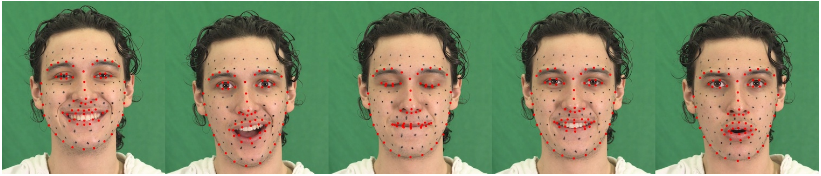
Point Distribution Models



Manually hand label a selection of images from a training set.

All examples *must* have the **same number** of landmarks and be labelled in the **same order**.

Point Distribution Models



Sufficient images must be labelled to capture the expected range of variation.

- Capture large facial expressions, wide mouths, etc.
- Typically need 20 - 30 images per person.

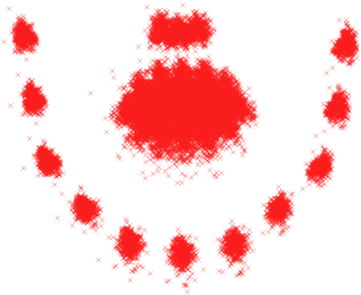
Point Distribution Models

A shape is the concatenation of the x and y coordinates of the landmarks:

$$X = \{x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n\}^T$$

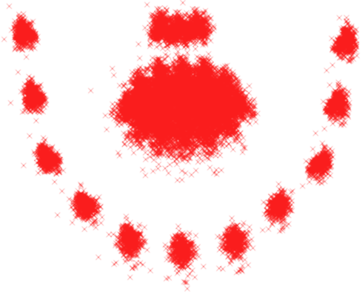
The consistency in the labelling ensures the elements of these vectors have the same meaning.

Point Distribution Models



The coordinates describe the shape in the image coordinate frame.
The same shape at different locations results in a different shape vector.

Point Distribution Models



We need to normalise shapes for translation, scale and rotation. This can be done using **Procrustes analysis**.

Aside: Procrustes analysis

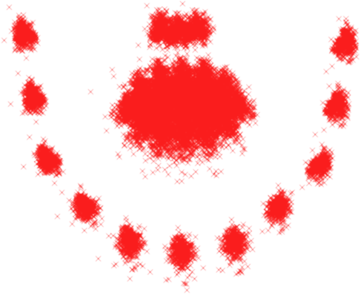


Figure 10: captured landmarks

Figure 11: aligned landmarks

Point Distribution Models

Given the aligned shapes, compute a model that describes the variation in shape.

A linear model of the variation can be found using **Principal Components Analysis (PCA)**.

Point Distribution Models

The model is in the form:

$$x = \overline{x} + \mathbf{P}_s \mathbf{b}_s$$

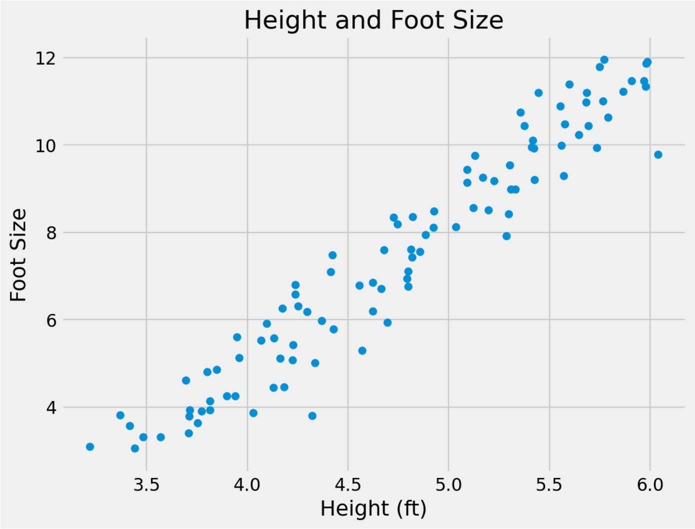
where x is a shape, $\overline{x}$ is the *mean* shape, the matrix $\mathbf{P}_s$ describes the variation in shape, and $\mathbf{b}_s$ are the **parameters** that represent a shape instance.

Aside: Principal Component Analysis (PCA)

- Reveals the internal structure of the data in a way that best *explains the variance* in the data.
- Used for dimensionality reduction.
- Reduces data down into its basic components, stripping away any unnecessary parts.

Aside: Principal Component Analysis (PCA)

- Assume we have 2-dimensional measurements. e.g. the height and foot size for a number of people
- We expect the measurements to be correlated to some degree. e.g. taller people tend to have larger feet
- Visualise the data by plotting one measure against the other.

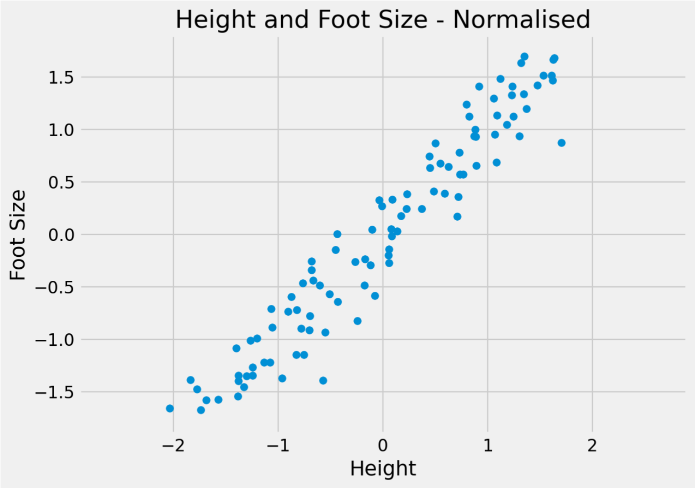

Height and Foot Size
Foot Size
Height (ft)
12
10
8
6
4
3.5
4.0
4.5
5.0
5.5
6.0

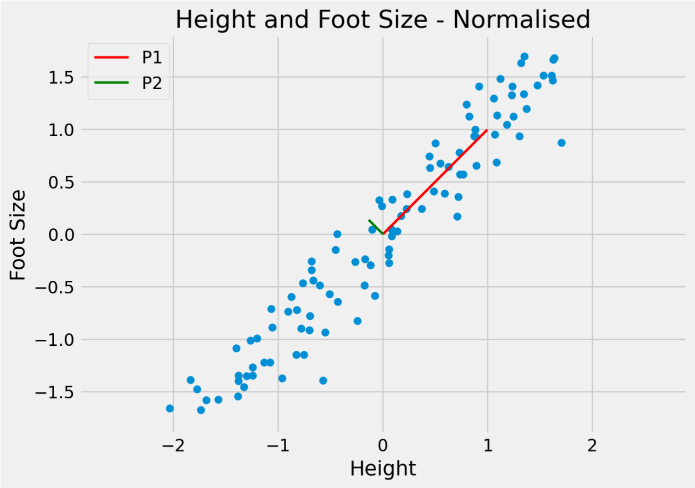
Aside: Principal Component Analysis (PCA)

The objective of PCA is to capture as much of the variation in as few dimensions as possible.

Find line of "best fit" through the data, then line of "next best fit" which is *orthogonal* to the first...

Repeat for however many dimensions your data has


Height and Foot Size - Normalised
Foot Size
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2
−1
0
1
2
Height


Height and Foot Size - Normalised
P1
P2
Foot Size
Height
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2
−1
0
1
2

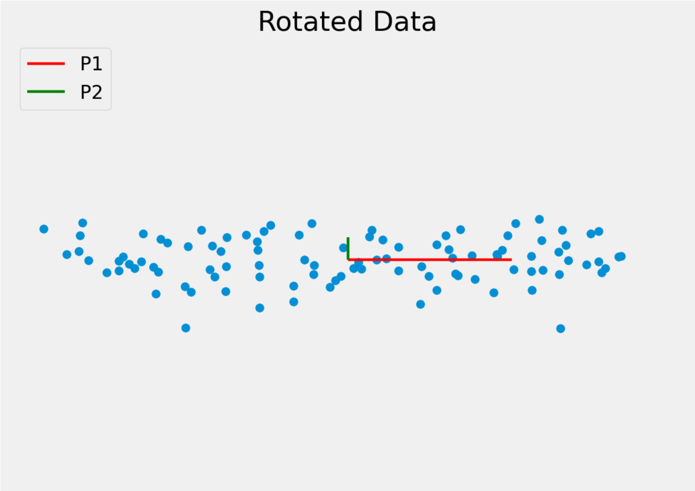
Aside: Principal Component Analysis (PCA)

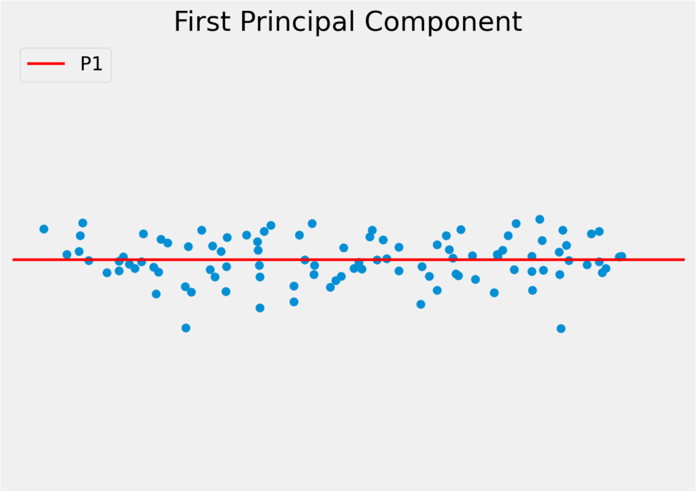
Since the dimensions must be orthogonal, all we have done is rotate the axes to better align with the data.

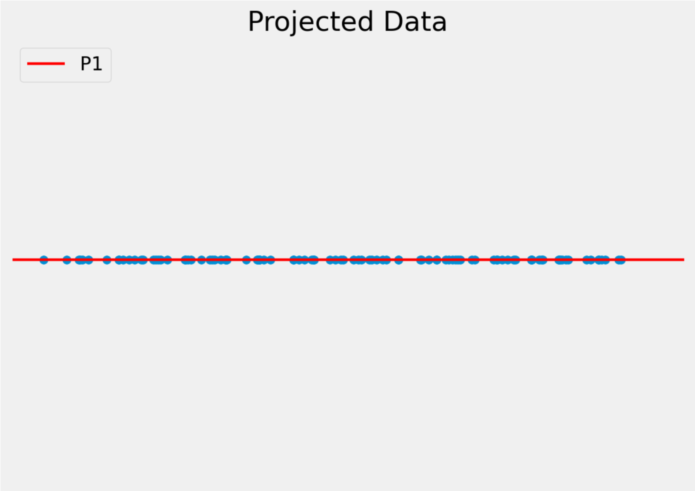
In doing this:

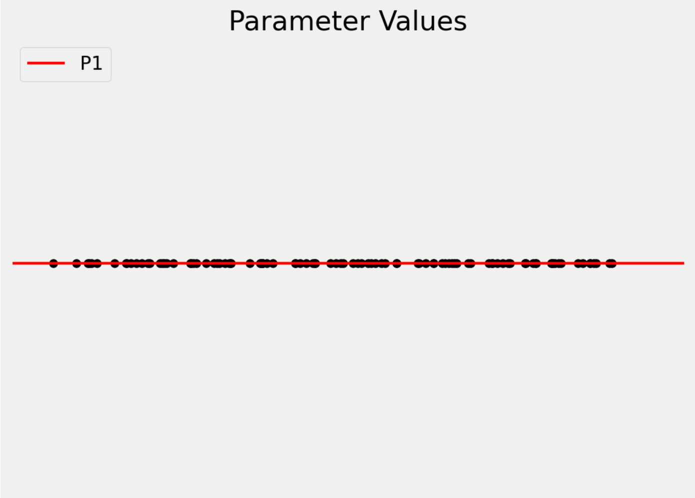
- P1 captures most of the meaningful variation
- P2 seems to capture the noise in the measurements

The original data can be approximated as some distance along P1 from the centre of the data cloud.


Rotated Data
P1
P2


First Principal Component
P1


Projected Data
P1


Parameter Values
P1

Aside: Principal Component Analysis (PCA)

To project a data point onto a new axis:

$$\mathbf{b}_s = \mathbf{P}_s^T (x - \overline{x})$$

Aside: Principal Component Analysis (PCA)

To reconstruct the data point from the features:

$$x \approx \overline{x} + \mathbf{P}_s \mathbf{b}_s$$

This is only an approximation since the data are truncated to lie on just the principal component(s).

Aside: Principal Component Analysis (PCA)

Note, in the previous example we have moved from a 2D problem to 1D so the representation is more compact.

Staying within the limits of the data means new examples can be generated — this is a **generative** model.

Aside: Principal Component Analysis (PCA)

Algorithm:

- Compute the mean of the data and subtract.
- Compute the covariance matrix.
- Compute the eigenvectors and eigenvalues of the covariance matrix and sort into descending order of eigenvalue.

Aside: Principal Component Analysis (PCA)

- Eigenvectors are the principal components.
- Eigenvalues are the variance explained by each principal component.
- We typically retain the number of eigenvectors that describe 95% of the total variation in the data.

Aside: Principal Component Analysis (PCA)

Matlab has implementations of both PCA and of Eigenvector/Eigenvalue decomposition.

Point Distribution Models

For modelling shapes, an n-point shape is represented as a 2n element vector:

$$X = \{x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n\}^T$$

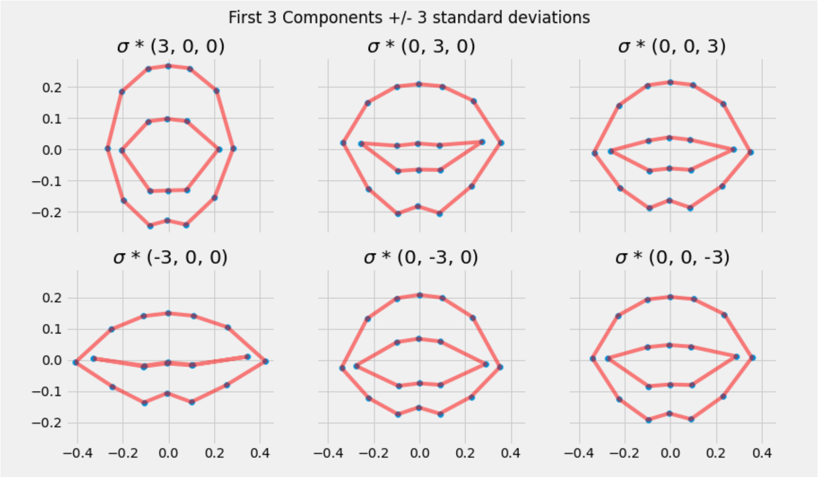
Can be thought of as a single point in a $\mathbb{R}^{2n}$ space.

Point Distribution Models

PCA can be applied to the $\mathbb{R}^{2n}$ data, rotating the $2n$ axes to best fit to the data cloud in $\mathbb{R}^{2n}$ space.

We retain only the meaningful variation - often resulting in considerable compression.

Point Distribution Models



Fitting a PDM

Given a PDM, and a new image, how do we fit the PDM to the facial pose in the new image?

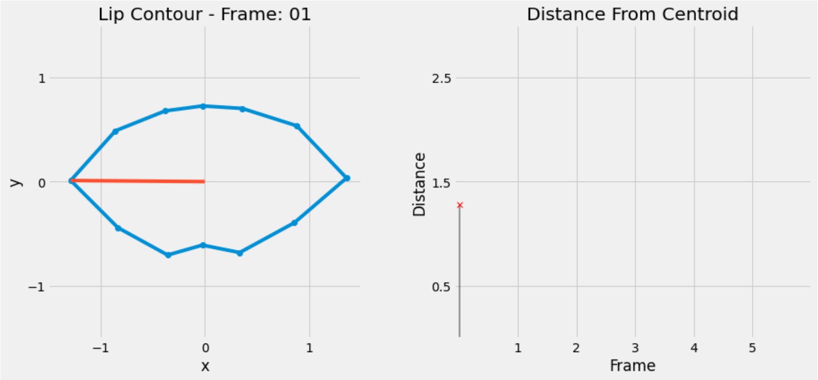
- Sample the pixels around each landmark in the training set, and look for the region in the image that best matches the sample.
- Refine the fit by forcing the shape to lie within the model space.
- More efficient if provided an approximate starting point.
- Further reading: Active Shape Models

Fourier Descriptors

The lip boundary provides a closed contour.

- Normalise the length to 2π units.
- Measure the distance from the centroid to the contour at regular intervals to calculate a **Centroid Contour Distance Curve**.
- The curve is *periodic* with period 2π, and it is real, continuous.

Fourier Descriptors



Fourier Descriptors

The curve can be decomposed into a **Fourier** series (refer back to the audio processing slides).

Fourier Descriptors

- The coefficients of the series provide the visual features
- This requires an accurate and complete estimate of the lip-contour.
- The coefficients do not have direct physical meaning.

Summary

- Visual Features
- Image segmentation
- Point distribution models and PCA
- Fourier descriptors