Visualisation and Topological Aspects of Higher Dimensional Data

Samuel Jackson
Department of Computer Science, Aberystwyth University

Introduction

Higher dimensional data is any data where the number of dimensions exceeds the three spatial dimensions we can interpret. Higher dimensional data has three major issues: increased processing time, the "curse of dimensionality", and visualisation complexity. This project aims to examine the effects of mapping a **high** dimensional space to a **low** dimensional one. Specifically, we are looking at mapping the higher dimensional **feature space** of both real and synthetic mammogram images to a lower dimensional representation.

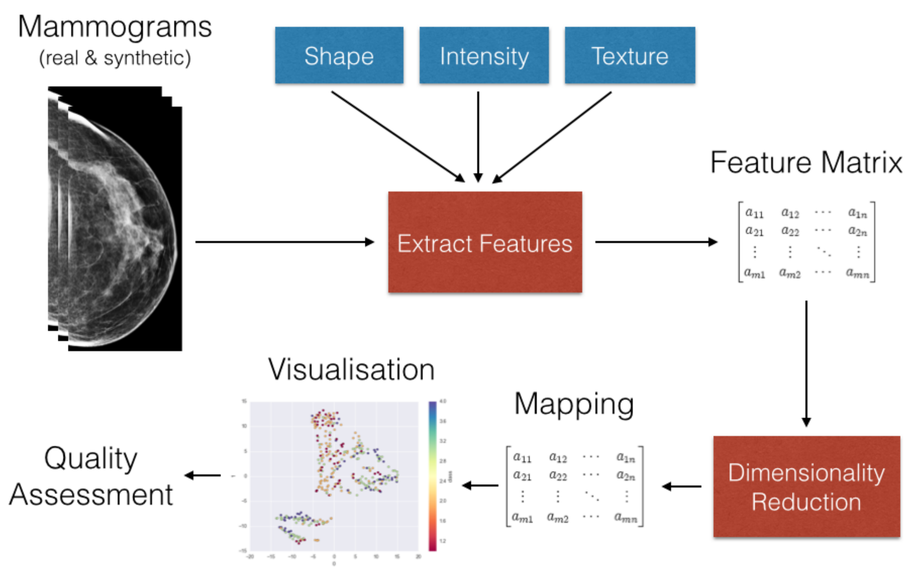
The goal of this study is to evaluate the mapping between the higher dimensional feature space and the lower dimensional representation.

- What is the relationship between the lower dimensional representation and the higher dimensional feature space?
- Does the synthetic dataset line up under the mapping?
- If not, does this match with the limitations discussed by the authors of the synthetic data?
- Could this information suggest how to build better models?
- Could this information influence a radiologists perception of what features are important in a mammogram?

Current Progress & Methods

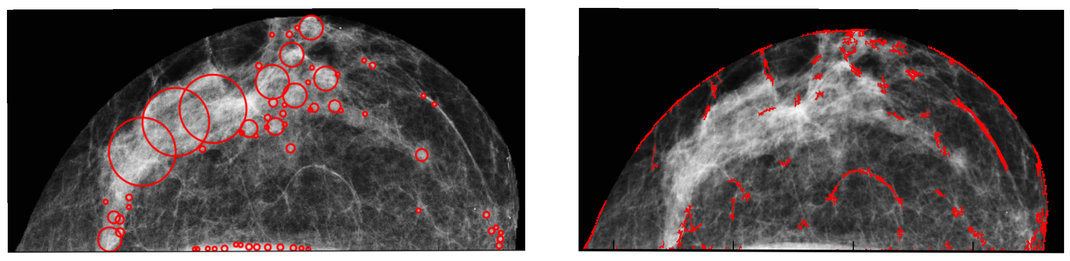
We are using a combination of shape, intensity, and texture image features. Regions of interest (ROIs) in a mammogram are detected using two types of shape features; blobs and linear structures. From these ROIs intensity and texture features are generated from the area of the ROI.

Once the feature set has been generated the results are run through a dimensionality reduction algorithm to produce a lower dimensional (2-3D) representation.



Shape Features

A multi-scale approach utilising a Laplacian of Gaussian pyramid is used to detect regions of interest (ROIs) corresponding to high density areas within the mammogram. Overlapping blobs are subsequently based on density and overlap [1].



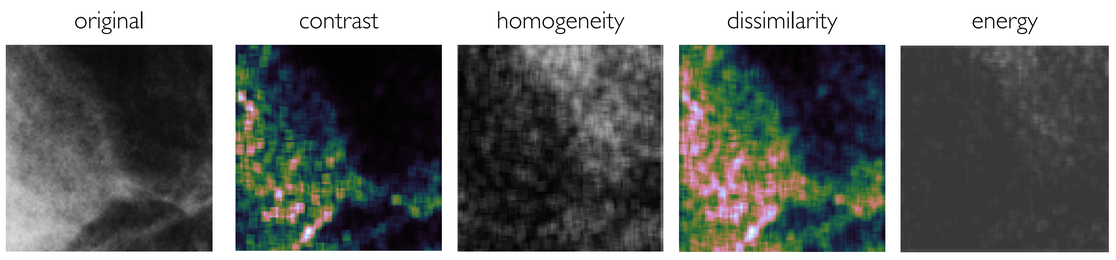
Linear structure is detected using an orientated bins feature. Non-maximal suppression and a Gaussian filter are used to improve response strength. Morphological closing is used to join close but segmented features [2].

Intensity Features

From the ROI identified by the shape features statistical features based on the intensity histogram of the ROI such as mean intensity, s.d. in intensity, lower and upper quartiles, skew, and kurtosis can be computed. These can later be grouped by scale.

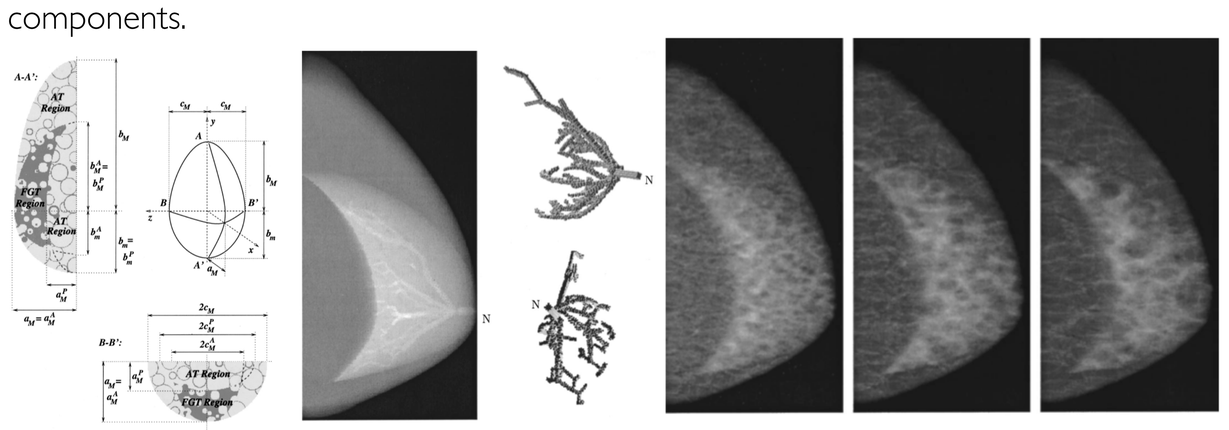
Texture Features

Texture features can also computed from ROIs detected using shape features. We use features based on the Grey Level Co-occurrence Matrix (GLCM) [4] which computes the number of occurrences that two neighbouring intensities appear next to one another. Features such as energy, homogeneity, contrast, and dissimilarity can then be generated from this.



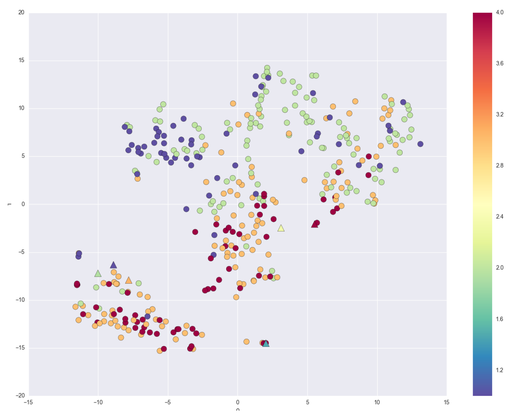
Breast Phantoms

The synthetic mammogram "phantoms" were generated from a 3D model outlined in ref. [3] by the University of Pennsylvania. The phantoms contain simulated blob and ductal structures inside a deformable model which is squashed like a real breast would be in order to produce a realistic looking mammogram in terms of structural components.

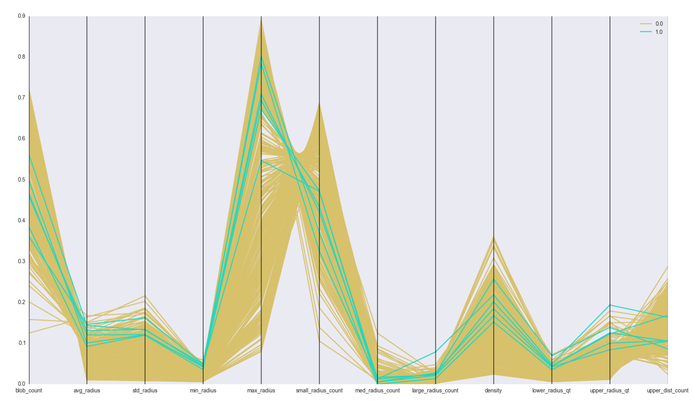


Results

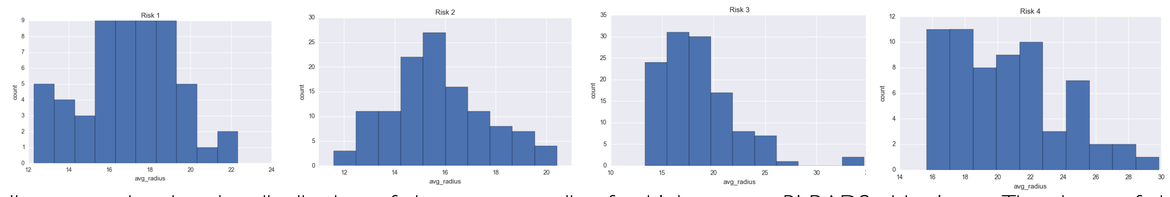
Dimensionality reduction is performed using the non-linear manifold learning algorithm t-distributed Stochastic Neighbour Embedding (t-SNE) [6] on each of the different feature spaces to generate lower dimensional mappings.



Mapping produced by t-SNE from features derived from the shape of blobs detected. Colours denote BI-RADS class.



Visualisation of the same feature space using parallel coordinates. Each axis represents a single dimension (feature). Green: phantoms, Yellow: real mammograms.



Histograms showing the distribution of the average radius for blobs across BI-RADS risk classes. The shape of these distributions and the distributions of other features allows the t-SNE algorithm to discriminate the most probable position of an image in the lower dimensional space.

Work Remaining

Line Feature Reduction

So far we have not performed a full run on the datasets using the line shape features. Like blob features, intensity and texture features will be computed for each of the ROIs selected by the lines.

Quality Assessment

Currently the evaluation of the mappings that have been produced have been evaluated based solely on the eye of the researcher. Further work needs to be done to automatically infer the quality of the mappings produced using measures such as trustworthiness and continuity [7].

Further Development

Other Techniques

In further work additional techniques could be explored, such as spectral based dimensionality reduction and additional feature extraction approaches, such as using Gabor filters for texture.

Topological Data Analysis

Apart from additional features, one aspect of the project that we are unlikely to have time for is an investigation into the underlying structure of the feature space using topological data analysis. This could utilise the mapper library and the techniques outlined in ref [5].

References

1. Chen, Zhili, et al. "A multiscale blob representation of mammographic parenchymal patterns and mammographic risk assessment." *Computer Analysis of Images and Patterns*. Springer Berlin Heidelberg, 2013.
2. Zwiggelaar, Reyer, Tim C. Parr, and Christopher J. Taylor. "Finding Orientated Line Patterns in Digital Mammographic Images." *BMVC*. 1996.
3. Bakic, Predrag R., et al. "Mammogram synthesis using a 3D simulation. I. Breast tissue model and image acquisition simulation." *Medical physics* 29.9 (2002): 2131-2139.
4. Haralick, Robert M., Karthikeyan Shanmugam, and Its' Hak Dinstein. "Textural features for image classification." *Systems, Man and Cybernetics, IEEE Transactions on* 6 (1973): 610-621.
5. Carlsson, Gunnar. "Topology and data." Bulletin of the American Mathematical Society 46.2 (2009): 255-308.
6. Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." *Journal of Machine Learning Research* 9.2579-2605 (2008): 85.
7. L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. "Dimensionality Reduction: A Comparative Review". *Tilburg University Technical Report*, TiCC-TR 2009-005, 2009

slj11@aber.ac.uk