Visualisation and Topological Aspects of Higher Dimensional Data

Report Name	Outline Project Specification
Author (User Id)	Samuel Jackson (slj11)
Supervisor (User Id)	Reyer Zwiggelaar (rrz)
Module	CS39440
Degree Scheme	G601 (Software Engineering)
Date	January 28, 2015
Revision	0.1
Status	Draft

1 Project description

This project is designed to examine the properties of mapping higher dimensional data onto a lower dimensional representation using manifold learning techniques [5]. More specifically the project will aim to provide a study of using dimensionality reduction techniques on both real [9] and synthetic [1–3] mammogram datasets to evaluate their correlation under the mapping.

The main goal of the project will be to produce a processing pipeline that loads and pre-processes sample mammograms from both real and synthetic datasets. Feature extraction methods can then be used to find relevant features within both the real and phantom mammograms simultaneously. Once features have been extracted dimensionality reduction techniques can be applied and the results visualised in a lower dimensional space. Visualisation of the results does not necessarily need to be limited to 2-3 dimensions [4]. It is hoped that a clear pattern will be found between both the synthetic and real data with the results from both datasets appearing close to each other in the lower dimensional representation.

The choice of manifold learning algorithm and feature extraction techniques are the main components under consideration for this project. This will require further background reading and research to evaluate the best candidates for each of these components. Ideally the project should focus on feature extraction methods which are commonly used and well understood in mammogram analysis. A variety of different manifold learning approaches can then be applied to the extracted features to reduce the dimensionality of the resulting data and select only those features which are most relevant. There will also need to be careful consideration regarding what metrics are used to evaluate the quality of the results.

The result of the project will provide an evaluation of the similarities and differences between the lower dimensional mappings of a synthetic and real dataset and try to distinguish whether the differences are in agreement with the limitations discussed by the authors of the synthetic data. In particular it would be interesting to see the how well the different classes of risk line up under the lower dimensional mapping of the two datasets and which features selected using the dimensionality reduction techniques for each dataset.

At this stage it would be beneficial to leave the choice of language and platform for the implementation open until further reading has been conducted. However, initial reading and discussion suggest that Python and/or C++ seem to be sensible candidate languages. Python in particular has lots of supporting libraries such as scikit-learn [8] and OpenCV [7] which will aid the implementation.

2 Proposed tasks

- **Background Reading & Research**: Initial research needs to be conducted to select appropriate techniques to be used for both the feature extraction and manifold learning components. Ideally this should include techniques commonly used by the mammogram analysis community (particularly in the case of features to be used) and should include research into how to compare the quality of the mappings. This should also include an evaluation of any previous, similar research that may have been conducted. A short-list of appropriate techniques should be chosen for use in the project's implementation.

- **Research into Implementation Technologies**: The choice of languages and technologies used in the implementation needs to be examined. As mentioned previously, on the surface Python appears to be an good choice, but a brief but more thorough study should be done.

- **Implementation**: The project needs to be implemented using the chosen technology and using the selected techniques. The implementation should include a test suite for each of the components in the pipeline as well as a way to visualise and compare the results.

- **Evaluation of Results**: Once the pipeline has been implemented, the results of using different feature extraction and manifold learning techniques on the real and artificial datasets should be examined and conclusions drawn from the experiment. These should be presented as part of the final dissertation report. Key questions for consideration are whether there is a correlation between the lower dimensional mapping of the real and synthetic data. If there is then is it what we should expect? If not, then how do they differ and is this to be expected based on the limitations of the synthetic model?

3 Project deliverables

- **Review of Research**: A report detailing the techniques for feature extraction, manifold learning, visualisation, and evaluation to be used in the system. This should include a justification of the choices made and will most likely also be included in the final report, but should be delivered earlier in the project before implementation begins.
- **Final Implementation**: A final implementation of the system should be produced as part of the project. This should provide a pipeline for transforming real & phantom mammograms to their representation as features and then map them to their lower dimensional representation and include a way to sensibly visualise the results.
- **Documentation of Final System**: There should be some documentation produced describing the system at a low level. This should include instructions for setting up and running the system and should include specifics about the implementation of each of the components in the pipeline.
- **Final Report**: A final dissertation report should be produced. This should include and overview of the techniques researched as part of the project, the techniques that were selected for the implementation in the pipeline (and why), a discussion of the system produced, a discussion of the results found, and an evaluation of the project itself.

Annotated Bibliography

[1] P. R. Bakic, M. Albert, D. Brzakovic, and A. D. Maidment, "Mammogram synthesis using a 3d simulation. I. breast tissue model and image acquisition simulation," *Medical physics*, vol. 29, no. 9, pp. 2131–2139, 2002.

A paper describing breast tissue modelling and acquisition process used by Bakic et. al.

[2] ——, "Mammogram synthesis using a 3d simulation. II. evaluation of synthetic mammogram texture," *Medical physics*, vol. 29, no. 9, pp. 2140–2151, 2002.

A second paper describing the simulated mammogram texture produced by Bakic et. al.

[3] ——, "Mammogram synthesis using a three-dimensional simulation. III. modeling and evaluation of the breast ductal network," *Medical physics*, vol. 30, no. 7, pp. 1914–1925, 2003.

A third paper describing the simulated breast ductal network produced by Bakic et. al.

[4] E. Bertini, A. Tatu, and D. Keim, "Quality metrics in high-dimensional data visualization: an overview and systematization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2203–2212, 2011.

Useful paper reviewing quality metrics and visualisation in regards to higher-dimensional data.

[5] L. Cayton, "Algorithms for manifold learning," *Univ. of California at San Diego Tech. Rep*, pp. 1–17, 2005.

Review paper on some common manifold learning algorithms. This is now likely to be a bit dated but provides a good introduction to the topic.

[6] K. Ganesan, U. Acharya, C. K. Chua, L. C. Min, K. Abraham, and K. Ng, "Computer-aided breast cancer detection using mammograms: a review," *Biomedical Engineering, IEEE Reviews in*, vol. 6, pp. 77–98, 2013.

Review paper on CAD with mammograms. This is both quite recent and provides fairly comprehensive overview of processing mammograms.

[7] OpenCV. (2015) OpenCV website. Accessed: 27/01/2015. [Online]. Available: http://opencv.org

Website for the OpenCV library

[8] Scikit-learn. (2015) Scikit-learn website. Accessed: 27/01/2015. [Online]. Available: http://scikit-learn.org/stable/

Website for the scikit-learn library

[9] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S.-L. Kok, *et al.*, "The mammographic image analysis society digital mammogram database," 1994.

Paper describing the Mammographic Image Analysis Society (MIAS) database