High-volume data processing for ambient healthcare research

Emma Tonkin

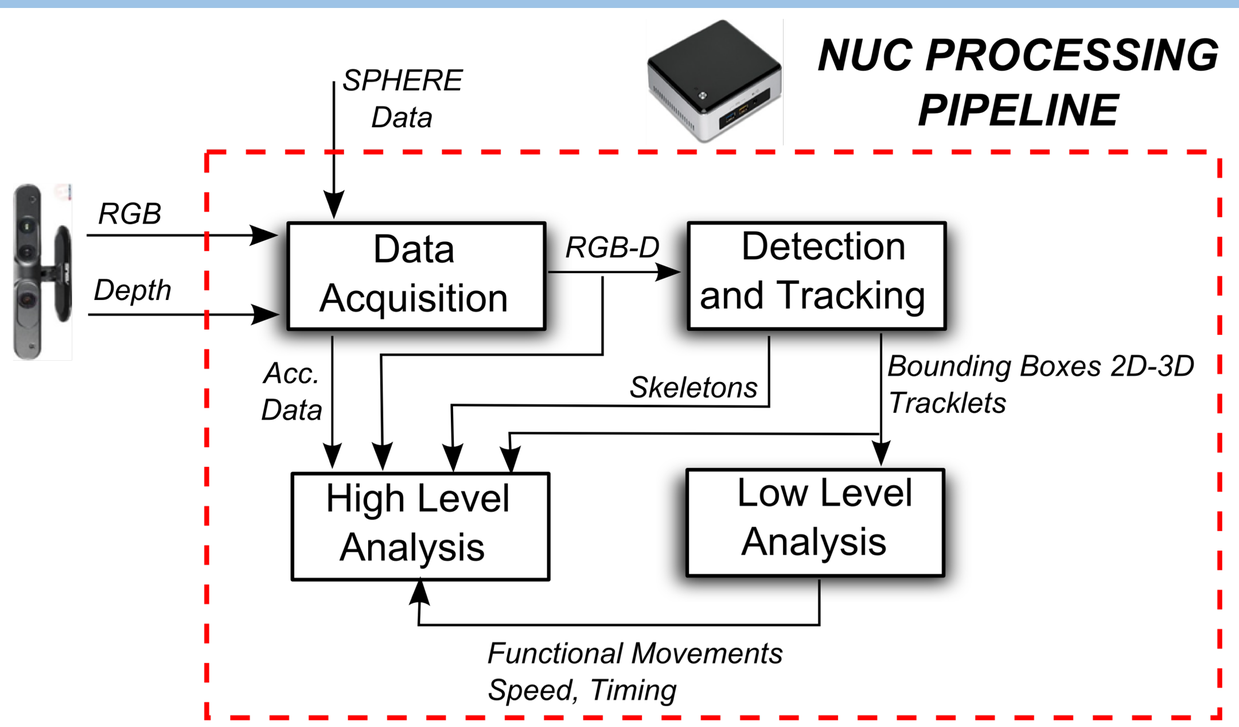
Faculty of Engineering, University of Bristol

ABSTRACT

The SPHERE project continues to deploy multimodal sensor networks into dozens of homes in the South West of England. The resulting dataset is expected ultimately to be large enough to significantly complicate reprocessing using the methods developed during the prototype phase of SPHERE, which were designed in the expectation that they would be deployed locally to each home. Therefore, various alternative means for processing the data are considered, including distributed indexing execution. For reasons of confidentiality and pragmatism, the use of public cloud platforms is not appropriate. The University of Bristol's HPC facility, Bluecrystal, will be used for this purpose.

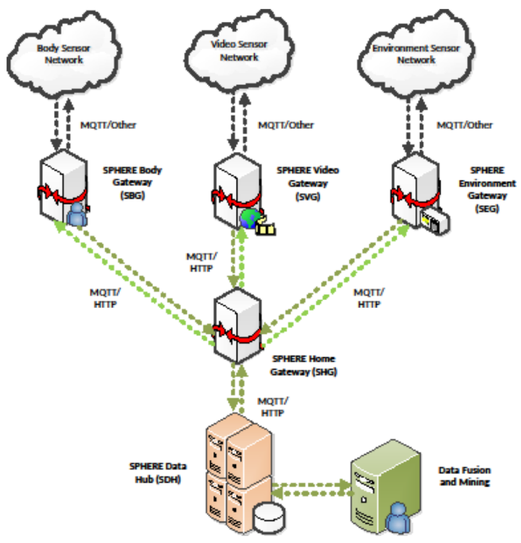
We are currently in the process of adapting our existing processing pipelines to the platform. Once this is complete, we expect the increase in processing speed and parallelism to permit us to compare the behaviour of several machine learning approaches across the dataset. Where practical, the most successful methods will eventually be adapted to run on the sensor platforms themselves.

SPHERE SENSOR NETWORK



Some analysis is completed on the distributed platforms used to host each individual sensor, as in the case of the video sensor, above. Both the sensor data and any initial analysis are then sent via MQTT to the Sphere Home Gateway, where it is further analysed using supervised machine learning methods trained in each deployed home. It is then securely stored on an encrypted drive. The decision to use physical medium transfer is taken due to the size of the dataset.

EXPLORATORY DATA FUSION AND MINING



The current processing pipelines for data within the home rely on a workflow-driven processing engine, Hyperstream. Currently the data processing depends on MongoDB, a NoSQL database, and preliminary post-processing was done through Python code, particularly SKLearn. This approach is well-suited to distributed deployment – it is run on the Sphere Home Gateway (see figure 2, left). However, it requires parallelization for effective deployment on the BlueCrystal cluster, and adaptations designed to take into account the architectural features of and constraints imposed by an MPI environment.

CONCLUSIONS

Moving our post-hoc exploratory data fusion and mining activities into the HPC environment is expected to provide us with the flexibility to increase the scope of our work beyond low numbers of individual exemplars and giving us an overview of performance across the full set of deployments. This work is currently ongoing, and is additionally expected to feed into a novel indexing method for data drawn from wearable and ubiquitous sensors.

University of BRISTOL

EPSRC Engineering and Physical Sciences Research Council

References

1. P. Woznowski, R. King, W. Harwin,and I. Craddock, "A human activity recognition framework for healthcare applications: ontology, labelling strategies, and best practice," in Proceedings of the International Conference on Internet of Things and Big Data (IoTBD), 2016, pp. 369–377.

Acknowledgements

This work was performed under the SPHERE IRC, funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/K031910/1. We would like to thank the participants of the study for their time and insights.