Chapter 6

Concluding Remarks

6.1 Summary and Conclusion

Current information technology development coupled to modern astronomical instrumentation is leading to accelerated acquisition of large volumes of observational data, fast pipeline processing and enormous on-line archives containing high quality images and multi-dimensional parameter values of astronomical objects. This is the case for facilities covering a wide spectral range. A few recent examples are the Sloan Digital Sky Survey (SDSS), the 2 Micron All Sky Survey (2MASS) and the NRAO VLA Sky Survey (NVSS). Each of these provide catalogues of many millions of objects, with often several tens of parameters characterizing the objects. In the near future these data volumes will increase by an order of magnitude when a new generation of instruments comes on-line.

In this thesis, we proposed visual and computational paradigms to analyze and extract information out of this flood of data. To obtain such techniques problems of a twofold nature needed to be overcome: one is the huge size of the datasets and the other is their large dimensionality. Density estimation approaches can be used for handling large size. There exist several techniques that can be used to obtain density profiles of the data. However, if we want to use the outcome of such estimators in later stages of the process, they need to fulfill certain criteria, such as computational efficiency, correctness, etc.

In Chapter 2 we studied the performance of four density estimation techniques: k-nearest neighbors (kNN), adaptive Gaussian kernel density estimation (DEDICA), a special case of adaptive Epanechnikov kernel density estimation (MBE), and the Delaunay tessellation field estimator (DTFE). The adaptive kernel based methods, especially MBE, perform better than the other methods in terms of calculating the density properly, and have stronger predictive power in astronomical use cases. Moreover, the computation time of these methods is lower than other methods and they compute the density on grids which can facilitate visualization (as an image in 2D and a volume in 3D) and analysis of the data. Using the MBE method we can also achieve scalability in terms of number of data points. After the original feature space has been transformed into image space, further computation can be done in image space that has constant size, although the size of the dataset can grow very large.

The next step is to extract useful information from such spaces. Clustering is one of the techniques that can help discovering structures in a dataset. However, full-dimensional clustering is not so useful since structures may exist in different subspaces that may indicate different relations among particular subsets of dimensions. Subspace clustering is an approach that can be applied for this purpose. Subspace clustering is the process of finding clusters in subspaces of the full feature space, either directly (Agrawal *et al.* 1998) or by identifying relevant subspaces for (later) clustering based on some quality criteria (Baumgartner *et al.* 2004). In Chapter 3 (Ferdosi *et al.* 2010), we proposed an interactive approach to find relevant subspaces which is strongly tied to morphological properties of object distributions. We used connected morphological operators implemented using the Max-tree data structure to identify the clusters (high-intensity regions in the density image). A "quality" of the subspaces was defined depending on their clustering property. We recovered various known relations directly from the data with little or no *a priori* assumptions. Therefore, our method can act as a starting point in analyzing large datasets and help users to find new relations as well.

Using the method described in Chapter 3, we can obtain interesting subspaces of any dimension. However, visualizing high-dimensional structures in a meaningful and user-interpretable way is far from straightforward. Traditionally, low-dimensional representations of high-dimensional spaces, obtained by methods such as Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), etc., are used to perform visualization in a Cartesian coordinate system. However, they pose the problem of interpretation of the visualization, because of the transformation of the original feature space to a new coordinate system. Two widely used methods to visualize high-dimensional data without transformation are the scatter plot matrix (SPM) and the parallel coordinate plot (PCP). For effective visualization of high-dimensional structures, they also requires a proper ordering of the dimensions. In Chapter 4 (Ferdosi and Roerdink 2011), we proposed algorithms for reordering dimensions in PCP and SPM in such a way that high-dimensional structures (if present) become easier to perceive. We used the quality criterion and the cluster indication capability of the method described in Chapter 3 to present three algorithms: two for finding suitable dimension ordering for PCP and one for SPM.

Algorithms presented in this thesis can be very helpful in visualizing and extracting information from large high-dimensional datasets, not only from astronomy but also from other domains where the datasets are of similar nature. Experimental results with synthetic and real life astronomical and gene expression datasets proved their potential.

In Chapter 5, we discussed several design issues of a visual analytic tool for astronomical data using a large touch sensitive display and present a prototype for such a tool. Large touch-sensitive displays provide more screen space, support more intuitive and natural interactions with touch sensitive inputs, and can make sharing of the analysis process and collaboration with others possible. Thus, they can facilitate analysis of astronomical data which are not only large in size and dimension but also complex in nature.

6.2 Future Outlook

6.2.1 Visual Clustering

The algorithms proposed in Chapter 3 and Chapter 4 retrieve clustering information without doing the clustering. However, it is possible to perform the actual clustering using the same method. In the proposed methods we only keep track of the pixel intensities of the connected components. It is also possible to identify the positions of the connected components with high intensity using the Max-tree. The pixel positions of the dense regions can be used to provide visual clues about the position of the clusters. Depending on the size of the connected component it may also possible to draw a circle centering at the pixel position of each cluster center. Further refinement of cluster selection can be done by user interaction. This type of visual clustering can be very helpful for datasets such as the Galactic stellar halo dataset used in Chapter 3 where traditional clustering methods such as k-means clustering produce unsatisfactory results.

6.2.2 Subspace Ranking for Outlier Detection

Outlier detection is one of the tasks that attracts a lot of attention from the astronomers. In a group of galaxies an outlier may indicate findings of a quasar[1]. In the literature, most of the time outliers are identified as a by-product of a clustering task. It is also possible to extend our subspace ranking for clustering to subspace ranking for outlier detection.

6.2.3 Analyzing Astronomical Datasets in Co-located Collaborative Environment

Collaboration, both intra- and inter-disciplinary, plays an important role in scientific research. Co-located collaboration has been practiced for a long time in scientific communities. With the advent of new technologies in communication, remote collaboration became quite common as well. However, increasing size and complexity of the datasets and new developments in display technologies, such as wall or tabletop displays, geared up the research on co-located collaboration using visual aids. To facilitate such collaboration there arises the need for collaborative visualization of data.

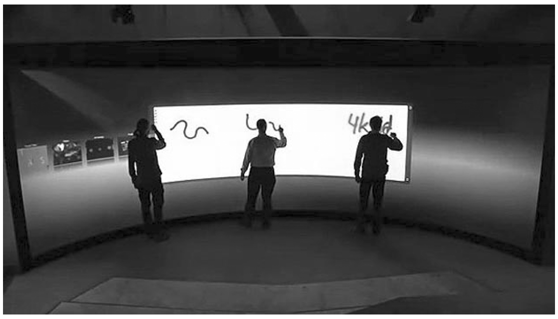

Figure 6.1. *Curved 396" wide touch-screen with 4900 x 1700 pixel resolution and 100+ simultaneous touches in the Reality Theatre of the University of Groningen.*

In Chapter 5 we discussed different design issues of a visual analytic tool in a two-touch wall display. It has the potential for being transferred into a co-located collaborative environment.

[1] Active galactic nucleus with very high redshift.

However, in that case we will need a touch-display with more than two-touch sensitivity. The Donald Smits Center for Information Technology of the University of Groningen owns a curved 396” wide touch-screen with a resolution of (4900 × 1700) pixels and 100+ simultaneous touch (see Figure 6.1). For designing a tool for co-located collaboration this screen may offer the best platform.

There are several studies in the literature trying to understand how co-located collaboration works in table-top displays. However, there is a lack of studies how such collaboration may work in large wall displays. Before designing tools in such environments it will be necessary to study how co-located collaboration may take place in such displays.