Scalable analysis and visualization of high-dimensional astronomical data sets
Ferdosi, Bilkis Jamal

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2011

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Chapter 5
Towards the Design of a Visual Analytics Tool for Astronomical Data on Large, Touch-Sensitive Displays

Abstract
Astronomical datasets are not only large in size and dimension but also complex in nature. To facilitate the analysis of astronomical data there exist several tools in the astronomical community. Although they provide a large range of visualization facilities, none of them provide any guidance for exploring high-dimensional datasets that have a large search space. The large size of the search space prohibits manual exploration and needs automation to some degree. Full automation is not desirable when the task is to find new or unknown phenomena. The use of visual analytics, where automation is coupled with user validation through interaction, can be useful. In this chapter, we take the first steps in designing a visual analytics tool for analyzing astronomical datasets on wall displays that integrates methods proposed in Chapter 3 with multi-touch interactions that provide 7DOF for spatial navigation. We derive several requirements for supporting visual analytics which in part are derived from extensive discussions with astronomers. We implement a prototype that fulfills the derived requirements and outline a plan for further refinement through user evaluation.

5.1 Introduction
Astronomical datasets such as catalogue data of astronomical objects (e.g., stars, galaxies) can consist of tens of terabytes of complex data. In near future, the size of the datasets will reach in the peta-byte range. To retrieve information from such information-rich datasets, automated data analysis tools may be required. However, a full automation may not sufficient when the task is to find new and unknown phenomena. Applying a visual analytics approach where computer automation is coupled with the power of the human brain in identifying structures, trends, or outliers (here, Chapter 1 of the book by Keim et al. (2010) can be useful). Visual analytics can make automatic analysis processes more transparent to the users because it supports visualization of the intermediate results. It also helps in better understanding of the obtained information as
it allows users to interact with the system and observe the effect of change in parameters on the end result. Therefore, using visual analytics approaches for analyzing astronomical data can be useful as well.

In Chapter 3, we presented a visual analytics approach to recover relevant subspaces for clustering. There, we visualized intermediate density images to facilitate the setting of smoothing parameters, and the end results were shown in an interface with a tree visualization of the ranked subspaces. The interface also integrated three other visualization tools (MTdemo, TOPCAT, and GGobi) to ease the further exploration of the result. However, using three separate interfaces for further analysis is time-consuming and can compromise consistency. For example, MTdemo visualizes 3D density images, TOPCAT provides 1D/2D/3D scatter plots, and GGobi has several information visualization techniques such as 1D/2D scatter plots, parallel coordinate plots, tours etc., with linked views. Among them, the performance of GGobi is not satisfactory if the dataset is large (> 50,000 data points). Interaction with the plots becomes very slow and structure identification in parallel coordinate plots becomes hard to achieve due to over-plotting of data.

To overcome these difficulties, it will be useful to integrate the required facilities and visualization techniques used in the system as proposed in Chapter 3 into a single interface. However, more screen space may be needed to design such an interface. An interface that provides different types of visualization and that facilitates the analysis of different types of data with different dimensionality may not be very useful in the case of regular desktops, because this may result in too much information in too small a space. Overcrowding of windows and clutter in plots can make the analysis difficult as well. Large touch-sensitive displays can be suitable for this purpose because of the large screen space they offer.

Large displays generally exist in two different orientations: horizontal (tabletops displays) and vertical (wall displays). Both types of display have advantages and disadvantages. Tabletops are well known for supporting co-located collaboration, e.g., Forlines and Shen (2005), Isenberg et al. (2010), Voida et al. (2009), Isenberg et al. (2009a), since they expedite different mechanics of collaboration such as monitoring, coordination of action, non-verbal communication such as eye contact and gesturing (Mandryk et al. 2002). For wall displays there is a lack of studies how such collaboration may work one such study can be found in Isenberg et al. 2009b, but they have the advantage that they provide a familiar display perspective to the user, with the same orientation as the common desktop displays.

In this chapter, we discuss several design issues of a visual analytics tool that facilitates analysis of astronomical data on a large wall display with touch-sensitive interactions and we present a prototype of such a tool (see Figure 5.1). The technique presented in Chapter 3 is used to support visual analytics of the datasets. For the touch interactions we use the technique presented by Yu et al. (2010) that provides 7 DOF (Degrees Of Freedom) for spatial navigation in 3D space.

The remainder of the chapter is organized...
as follows. In Section 5.2 we discuss related work. A brief description of the necessary background is provided in Section 5.3. Requirements of the tool and a description of the proposed interface are presented in Section 5.4. We conclude the chapter in Section 5.5 along with plans for future work.

5.2 Related Work

The proposed tool is designed to work with astronomical data that possesses features from visual analytics, scientific visualization, and information visualization. Therefore, we discuss related work in four different groups: the first group concerns several existing astronomical tools, the second group some existing information visualization tools, the third group concerns some domain-specific tools that integrate scientific and information visualization, and in the fourth group we discuss some visual analytics tools that use large touch displays. We discuss their strengths and shortcomings and try to indicate how the proposed tool can improve upon them.

A detailed discussion of existing astronomical tools can be found in Chapter 1. In this section, we summarize their strengths and shortcomings. Most of the existing astronomical tools are designed to address different astronomical data analysis and visualization needs. TOPCAT is a powerful tool for working with large tables but lacks data visualization options. Aladin is useful for image-related operations but not well suited for data analysis purposes. VisIVO and VisIt offer several options for visualization, but they do not provide any guidance to the user to initiate the analysis process, especially when the number of search spaces is very large due to high data dimensionality. These tools can be useful when the user knows where to look for certain information. However, in such cases a huge amount of information hidden in other subspaces may remain unnoticed. Information visualization tools such as GGobi and XmdvTool provide extensive information visualization facilities. They also provide on-demand visualization and are suited for relatively low-dimensional datasets where manual exploration of information is still possible. Unlike these tools, in our proposed tool we provide a ranking of subspaces for clustering, so that users can start the analysis based on the best views of the dataset. It also allows users to change parameters interactively and observe the respective changes in the ranking of the subspaces.

There is a large number of domain-specific systems that integrate 2D/3D scientific visualization with information visualization, such as systems for DTI (Diffusion Tensor Imaging) fiber exploration (Chen et al. 2009, Jianu et al. 2009), or geographic information systems, such as GRASS (Neteler and Mitasova 2002), SAGA (Böhner et al. 2006), etc. There exist some visual analytics systems that also use large displays (Schreck et al. 2007, Tekušová and Kohlhammer 2008). Among these, Schreck et al. (2007) presented a system for analyzing large financial time series data using trajectory-based visualization. Another visual analytics tool for exploring and analyzing complex corporate shareholder networks is proposed by Tekušová and Kohlhammer (2008). Both of these tools use the IGD-HEyeWall of the Fraunhofer-Institut für Graphische Datenverarbeitung (IGD) in Darmstadt, Germany. These domain-specific systems meet the need for analyzing data from the specific domains mentioned, but are not suitable for use in other domains such as astronomy, since astronomy also has domain-specific requirements
that need to be addressed.

Therefore, our contribution is the design and implementation of a visual analytics tool for analyzing large (in size and dimension) astronomical datasets in a large touch-sensitive display. Unlike other existing tools our approach

- considers the domain-specific needs of astronomical data such as spatial and non-spatial visualization of astronomical objects, exploration of time series data, large scale and precise interaction with the astronomical objects etc.,
- allows users to find relevant subspaces for clustering high-dimensional data, and
- allows users to interact with the system to control the automatic subspace searching process to achieve the desired result.

5.3 Background

The proposed tool uses the subspace ranking for clustering technique to support visual analytics of large datasets. For interaction with the plots we use the frame-based interaction proposed by Yu et al. (2010). In this section, a brief discussion of the used approaches and the implications of these methods for our design decisions will be presented.

5.3.1 Subspace Ranking for Clustering

For the datasets with high dimensionality, the number of search spaces can increase very rapidly as a function of the number of dimensions. In a \(d\)-dimensional dataset the number of \(k\)-dimensional subspaces is \(\binom{d}{k}\), \(1 \leq k \leq d\), so the number of all possible subspaces is \(\sum_{k=1}^{d} \binom{d}{k} = 2^d - 1\). For example, a 3D dataset comprises 7 subspaces whereas a 10D dataset already contains 1,023 subspaces to search for information. Therefore, it requires automation to some extent to retrieve informative dimensions from this huge search space that prohibits manual search. However, full automation is not well suited for exploratory tasks. By finding the right balance between data mining approaches such as clustering on the one side and interactive visualization on the other we can enable users to explore such large numbers of subspaces more effectively.

The algorithm to obtain a ranking of the subspaces for clustering is based on a quality criterion. The quality of a subspace depends on the structure present in the data. Emphasis is given to multimodality of the density distribution of the subspaces, where each density mode is indicative of a cluster. In addition, the significance and separability of each mode contribute to the quality value. The search for the density modes and determination of significance and separability is performed in grey-level image space. Therefore, a transformation of parametric space to image space is required. This transformation can be obtained by grid-based density estimation. Thus, modes in the distribution are transformed into high-intensity peaks (local maxima) in the density image. Then we use connected morphological operators implemented via the Max-tree data structure to obtain the number of modes (counting the number of leaves in the Max-tree) and
their relative dynamics (see Section 3.3.4). Using this information we compute the quality of each subspace and rank the subspaces according to quality.

However, the choice of a proper smoothing parameter for creating the density image plays a vital role in the subspace searching process. If the smoothing parameter (window size of the kernel density estimator) is too large then an over-smoothed density image will be obtained and thus only large clusters can be identified. On the other hand, a too small smoothing parameter will result in an under-smoothed image where each data point is identified as a cluster. However, identifying over- or under-smoothed images automatically is not a trivial task for a computer, while human experts can easily identify such properties by observing the images. The choice of the smoothing parameter is also task-dependent. Depending on the task at hand the user may need to adjust the smoothing parameter in different ways. In Chapter 3 we provided an automatic choice for the smoothing parameter and then visualized the resulting density image for user validation. After observing the visualized image the user can then change the parameter setting, if necessary.

Therefore, to facilitate the interactive and iterative approach of parameter selection the interactive data exploration tool should provide a widget for changing parameter values and a dedicated visualization window. However, there should be two such interfaces: one for 2D subspaces, and another one for 3D subspaces or for the first three principal components of the subspaces with dimensionality higher than three. In addition, the tool should provide a way to navigate the ranked subspaces individually and also in groups to provide an overview of the result.

5.3.2 Touch Interactions with 7 DOF

Visualization and interaction go hand in hand. Visualization without useful interaction techniques is of little use. In our system, to explore the ranked subspaces for further analysis we used the FI3D frame-based interaction system proposed by Yu et al. (2010). This interaction system allows the manipulation of 3D space as a unit with 7 DOF (translation in x-, y-, and z-direction, orientation with respect to the 3D coordinate system, and uniform zoom) using only a single touch. In addition, the frame interaction widget also provides 2-touch interactions for RST (combined z-rotation, scaling, and x-/y-translation), and for constrained x-/y-/z-rotation. In Figure 5.2 the frame-based interactions are depicted. The four corners of the frame are dedicated to zoom in/out operations. Rotation around z-axis can be initiated by touching any side of the frame with an initial movement parallel to the frame. Rotation around the x-/y-axis is initiated by touching the frame with an initial movement perpendicular to the frame. Touching the center of the display area initiates x/y- translation. For the z-translation (moving the camera closer to/away from the dataset) two extra frame elements at the top and bottom of the frame are used. Touching any of these regions can initiate z-translation. All of these interaction regions serve to specify the interactions in spring-loaded mode fashion (Buxton 1986).

Integration of this interaction technique into our tool will enable both large-scale and precise interactions not only with the spatial 3D data but also with the non-spatial subspace plots and scatterplot diagrams. The FI3D interaction technique in its basic form requires just one touch-point. However, the technique can be easily extended to multiple touches.
Figure 5.2. FI3D touch interaction with 7DOF according to Yu et al. (2010)
Figure 5.3. Interface of proposed tool. Top: Buttons for individual exploration of spaces. Middle: (left) window for a (time dependent) 3D spatial plot; (middle) window for a 2D scatter plot / density plot; (right) 3D scatter plot / density plot. Bottom: parallel coordinate plot for parametric space.
5.4 Design Issues and the Prototype

In Section 5.3.1 we already mentioned some of the requirements that the proposed tool should fulfill to enable visual analytics of the data. Those requirements basically concern the parametric space of the dataset. However, in astronomy, spatial positions of astronomical objects play an important role in the analysis process. Astronomers are used to associate findings in parametric space with those in the spatial domain. Some astronomical data such as cosmological simulation data usually involve time-dependent spatial and non-spatial parameters. From discussions with the astronomers who work with the time-dependent datasets, we learned that tracing of selected particles in time is also of much interest in the analysis of such datasets. Therefore, the proposed tool should provide a dedicated interface for the spatial plots which is linked with the parametric space and which enables the exploration of time-dependent data.

The visualization requirements stated in Section 5.3.1 concern only 2D and 3D visualization, where subspaces of dimension higher than three are visualized after a transformation using principal component analysis. However, such transformations can cause a reduction of information as well. Therefore, visualizing these subspaces without any transformation in their original dimensionality can add value to the analysis. Thus, the tool should support high-dimensional data visualization as well.

In astronomical data analysis, such as the study of large scale structure of the universe, the interactions that allow both large scale and precise interactions with the data will be invaluable. Such interactions can be useful for other type of datasets as well.

To summarize the requirements stated above, the system should

\begin{enumerate}
\item \textbf{R1:} support simultaneous interactive exploration of spatial and non-spatial data,
\item \textbf{R2:} support visualization of time-dependent data,
\item \textbf{R3:} be able to trace selected particles in time,
\item \textbf{R4:} support visualization of data in 2D / 3D and higher than three dimensions,
\item \textbf{R5:} support visual analytics to find and explore large parametric spaces for finding trends / clusters,
\item \textbf{R6:} allow the user to control the findings by setting (smoothing) parameters,
\item \textbf{R7:} support linked views, and
\item \textbf{R8:} allow both large-scale and precise interactions.
\end{enumerate}

The visual analytics of the data should be provided for all of the dimensions. This is a bottom-up process, ranging from one-dimensional subspaces to \((d - 1)\)-dimensional subspaces, where \(d\) is the dimensionality of the dataset. However, it should be possible to obtain a ranking of a subspaces of any user-defined dimension. Also, the user may want to know the ranking of the subspaces that correspond to a user-defined feature and dimensionality. Therefore, the requirement \textbf{R4} can be divided into two sub-requirements:

\begin{enumerate}
\item \textbf{R5a:} support ranking of all \(d\)-dimensional subspaces and
\item \textbf{R5b:} support ranking of the subspaces of any user-defined dimension and/or with a user-chosen feature.
\end{enumerate}
Our tool intends to support visual analytics to find subspaces for clustering. However, usual scatter plot visualization is not sufficient to observe the clustering of the subspaces. Visualization of density images with color mapping can help in identifying structures. That leads us to another requirement of the system:

**R9:** provide visualization that emphasis the structures present in space.

The size and dimensionality of astronomical data can vary and can be very large, thus the tool should

**R10:** be scalable in terms of size and dimensionality of the dataset.

![Figure 5.4](image-url)  
*Top: Multi-purpose buttons. Bottom: Window for 2D subspace exploration with ranked subspaces on top.*

We developed a prototype for a 52 inch LCD screen with full HD resolution (1920×1080 pixels, 115.4 cm×64.5 cm). The display is equipped with a DViT overlay (Smart Technologies
Inc. 2003) from Smart Technologies, capable of recognizing two independent inputs. To fulfill requirement $R1$, we added four widgets: one for 3D spatial space and three for parametric space. Among the three widgets for parametric space, one is for 2D data, another is for 3D data, where both widgets offer visualization of data as scatter plot or density plot. The third widget is for the subspaces with dimensionality higher than three, visualized as a parallel coordinate plot, in this way fulfilling requirement $R4$.

To support time-dependent data navigation ($R2$) we added buttons at the bottom of the spatial window. Pressing on the right button increases the time step continuously and pressing on the left button decreases the time step. The middle button can be used to stop at any specific time step for analysis. Depending on the time step currently displayed in the spatial space, the corresponding parametric space visualization is updated as well.

Requirement $R5a$ is achieved with a button on the top right of the interface (see Figure 5.4) and the user can initiate $d$-dimensional subspace ranking by clicking on this button. This process uses the 2D and 3D windows to visualize the subspaces for user validation of the smoothing parameter. To allow the user to set the proper smoothing parameter ($R6$) we designed a touch-controllable slider next to the 2D and 3D windows. The user can observe the immediate change in the density image on the corresponding window whenever she changes the parameter by moving the slider. In order to navigate the ranked subspaces the user can click the buttons on the bottom of the corresponding window.

To achieve requirement $R5b$ we also use buttons, putting them in the top of the interface. By clicking on these buttons the user can initiate subspace ranking in any of the chosen dimensions. In this view, the user can also obtain an overview of the ranked subspaces (see Figure 5.4) and can observe the changes in the ranking with the change in smoothing parameter.

To emphasize the structures present in the data ($R8$) we use a color mapping of the 2D density plot initiated by the user’s demand. For the 3D case a color mapping is not useful as it hides the structures inside the volume, thus an X-ray rendering of the 3D density will be sufficient. To emphasize structures present in the data using parallel coordinate plots, we render it using histogram-based color coding.

The scalability issue ($R9$) has been addressed in the visual analytics part of the tool using image-based subspace searching (for a detailed description, see Chapter 3).

### 5.5 Summary and Conclusion

In this chapter, we have considered several design issues of a visual analytics tool for astronomical data and presented a sketch of a prototype. The main purpose of designing such a tool is to facilitate exploration and analysis of datasets with a large number of dimensions when the task is to find new or unknown phenomena. In addition, we provided an interaction approach that enables both large-scale and precise interactions with 7DOF. We designed the tool to be used with large touch-sensitive displays that provide more space for the visualization and analysis process. All these features make the tool different from other existing astronomical tools. The design decisions were made so that the tool will fulfill the requirements we derived to facilitate visual analytics of high-dimensional data. Many of these design decisions evolved from discussions
with astronomers.

Most of the design decisions to satisfy the desired requirements are already considered and implemented to some extent. However, the tool needs to go through user evaluation to validate the design decisions and refine them according to feedback from the users. Thus, future work will involve performing repeated user evaluations to refine all the design decisions. We also need to address the scalability issue regarding the visualization. Currently, for large datasets (e.g., with 3 million data points) we are able to visualize only 10% of the data (chosen randomly). The proposed tool also has the potential to be used in co-located collaborative environment. We will also consider this possibility in the future.

Acknowledgments

We thank dr. Tobias Isenberg for his valuable comments and suggestions.