Mem. S.A.It. Suppl. Vol. 13, 29
© SAIt 2009

Memorie della
Supplementi

VisIVO

Data exploration

M. Comparato and U. Becciani

Istituto Nazionale di Astrofisica – Osservatorio Astrofisico di Catania, Via Santa Sofia 78, I-95123 Catania, Italy, e-mail: marco.comparato@oact.inaf.it

Abstract. In this paper we talk about VisIVO, a novel open source graphics application, which blends high performance multidimensional visualization techniques and exploits the standards defined by the International Virtual Observatory Alliance in order to make it interoperable with VO data repositories. The paper describes the basic technical details and features of the software.

Key words. Data Analysis and Techniques

1. Introduction

The astronomical community has always dedicated special attention to the growth of graphical and visualization tools, driving their evolution or even being directly involved in the development of many of them.

At present, the most popular software for astronomers can be subdivided into two main categories: tools for image display and processing and tools for plotting data. Notable among the former are IRAF, by NOAO; ESO-MIDAS, by the European Southern Observatory; SaoImage, by the Smithsonian Astrophysical Observatory and GAIA, by ESO. Many other tools are available, but we refer to dedicated surveys for a complete list. Gnuplot and SuperMongo are popular applications adopted for 2D data plots. A more sophisticated solution is represented by IDL, by ITT Visual Information, which is characterized by a large library of functions specifically developed for astrophysics. Again, for a complete list, we refer to specific surveys.

Among the most popular N-body visualization codes used by the community there are: TIPSY, motivated by the need to quickly display and analyze the results of N-body simulations, it is mainly limited to this type of data; ParaView, produced by Kitware in conjunction with the Advanced Computing Laboratory at Los Alamos National Laboratory (LANL), the goal of the project is to develop scalable parallel processing tools with an emphasis on distributed memory implementations; IDL, mentioned above, contains support for N-body data display, but is not free software.

A new generation of graphic software tools is now emerging. These tools are designed to overcome the limits and the barriers of traditional software by exploiting the latest technological opportunities. The main challenges and objectives are the following:

– High performance and multi threading, in order to exploit multi-core systems, large

Send offprint requests to: M. Comparato memories and powerful graphic cards and co-processors. This allows the user to handle large amount of data in real-time.

– Interoperability, allowing different applications, each specialized in doing different things, to interact with each other in a coordinated and effective way according to well-defined protocols. The aim is to provide to the user a complete suite of tools to best analyze his/her data. Huge, monolithic and often inefficient tools are obsolete.
– Collaborative work. The tools allow several users to work on the same data at the same time from different places, exchanging experience, information and expertise.
– Access to distributed resources, via web services and/or Grid protocols. Often, data can no longer be moved from data centres as it is too large and complex. The astronomer must have the tools to access it, independently of his geographical location in a fast and reliable way.

Tools like VisIVO, Aladin (Bonnarel 2000) and Topcat (Topcat 2000), have been recently developed in the framework of the Virtual Observatory (http://www.ivoa.net) to achieve all or some of these goals. In this paper we will focus in particular on VisIVO, which stands for Visualization Interface for the Virtual Observatory. VisIVO is being developed as a collaboration between the Italian National Institute for Astrophysiscs (INAF) - Astrophysical Observatory of Catania and CINECA (the largest italian academic high performance computing centre) in the framework of the FP6 EU funded VO-Tech project. The next section gives a short review of the basic functionality of VisIVO.

2. VisIVO

VisIVO is a C++ application specifically designed to deal with multidimensional data. It is Free Software available both for MS Windows and for GNU/Linux (porting to MacOS is in progress). It can be downloaded from the web site http://visivo.oact.inaf.it. The software is built on the top of the Multimod Application Framework (MAF)(Viceconti 2004). MAF is an open source framework for the development of data visualization and analysis applications. It provides high level components that can be easily combined to develop a vertical application. It is being developed by the visualization group of CINECA and it can be downloaded from the web site http://openmaf.cineca.it. The framework is based on the Visualization ToolKit (VTK) (Schroeder 2004) library for the multidimensional visualization and on the wxWidgets library, a portable Open Source GUI library, for the user interface. VisIVO's architecture strictly reflects the structure of a typical scientific application built with the MAF, being mainly developed in the highest layers of the framework. The software exploits, wherever possible, the standard visualization services, views, operations and interface structures provided by the framework and implements all the elements that are specific to the visualization and analysis of astronomical data.

Extensions to the basic MAF infrastructure have been developed in order to match astronomy-specific requirements and to provide the highest performance. Internal data representation is in the form of a *Table Data* structure, which is composed of a sequence of variables loaded from a data source such as a file or a database. Regardless their original type, variables are all converted to *float* format. Once a table is loaded the user can manage and visualize the data. These operations do not increase the memory usage as long as they do not create new tables or new fields: the visualization process is carried out using references to the Table Data with no data replication. In order to visualize data, the user has to set which of the loaded fields will be used as the coordinate system of a Cartesian reference frame. In this way, the software ensures maximum flexibility in data usage.

2.1. VisIVO for data visualization

Data visualization is the main target of VisIVO. The software is designed to simultaneously handle as many properties as possible. Complex tables can be loaded and manipulated, new fields can be derived and finally represented graphically, using points, colours, transparencies, surfaces, glyphs and volume rendering. The first step of a working session is usually data loading. Data can be read from files; VisIVO supports different kinds of file formats: standard file formats, like VOTables, FITS, HDF5, ASCII, raw binaries and the native data format of the popular *Gadget* simulation code (Springel 2000). The VOTable format is an XML standard for the interchange of astronomical data, defined by the International Virtual Observatory Alliance (IVOA, http://www.ivoa.net). Data is represented as a set of tables, each table being an unordered set of rows, whose format is specified in the table XML metadata. Rows are sequences of table cells, each containing either a primitive data type or an array of such primitives. VOTables can also contain links to external files as a separate data source. FITS and HDF5 importers are implemented using the published API and libraries. The ASCII table format consists of columns of data spaced with the most common separation characters (space, tab etc.). Raw files are sequences of variables written as binary dumps of the memory. The binary files can be managed by descriptor files which store the associated information (number of variables, data types etc.). VisIVO can also interact with CDS VizieR data service (Ochsenbein 2000), retrieving data directly from remote archives (see par. 2.3).

Once data is loaded it can be visualized and analyzed. VisIVO can deal with both structured and unstructured data. It does not associate any geometry to the data, it is the user that, applying specific operations, can create geometries using one or many of the loaded fields. In this way, as examples, he can create bidimensional images using one field and specifying x and y resolution in agreement with the field dimension; he can create tridimensional grids using one field and specifying x, y and z resolution; he can create point distributions selecting three fields; and so on providing maximum flexibility.

After the user creates a point distribution using his data, points, besides their geometric position, can be used to display further quantities, using colours and glyphs (3D shapes, like spheres or cubes). Points can be coloured as a function of a given scalar field (e.g. their temperature or their spectral index) with a colourmap that the user can customize. Each point can also have an associated glyph, whose size can be a function of one (for spheres) or two (for cubes, cylinders, pyramids) fields. A vector quantity can be visualized as well, using either oriented segments or arrows. Vectors can also be coloured according to their magnitude.

If the data size is too large to be managed in memory, VisIVO allows the user also to extract a random subset of points. It is also possible to select the points which lie in a region the user is specifically interested in (e.g a galaxy cluster in a cosmological simulation, or a globular cluster in a catalogue of stars). The selection can be accomplished using either a rectangular sampler or the cluster finder utilities. For the latter, the following cluster identification method is implemented. A field associated to the points (e.g. the point mass) is used to set a threshold. All the points that have a value of the field above the threshold, comprise a cluster. Surfaces which divide regions above and below the threshold can be visualized (see figure 1). Regions geometrically disjoint (i.e. their threshold surfaces do not overlap) are identified as separate clusters.

Structured mesh-based data can be visualized using volume rendering and isosurface. Volume rendering is a visualization technique in which the field values are represented by different colours and different transparencies. The global effect is a cloud appearance. This method enables the user to emphasize also the inner parts of the volume. Isosurfaces are surfaces of given value calculated from a mesh-based quantity. The isosurface can be defined as the surface which divides regions in which a given field has value above a certain value from regions in which it is below that value.

2.2. VisIVO for data analysis

VisIVO provides various built-in utilities that allow the user to perform mathematical operations and to analyze data. It is possible to apply algebraic and mathematical operators to the

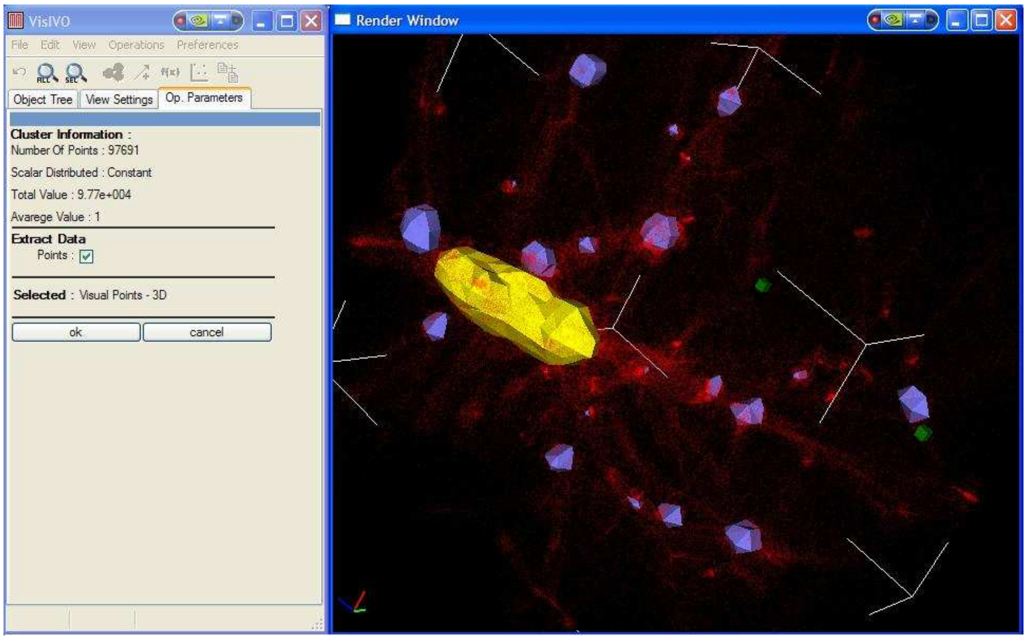

Fig. 1. Clusters identification in data from a cosmological simulations. Points inside the yellow isosurface can be extracted.

loaded data. Basic arithmetic operations (addition, subtraction, multiplication, division) as well as logarithm, power law, absolute value and many others are supported. Scalar product, magnitude and norm of vector quantities are available too. In this way, new physical quantities can be calculated. For example, for the gas distribution of a simulated galaxy cluster, the X-ray emission due to thermal Bremsstrahlung can be calculated as proportional to the product of the square of the mass density and the square root of the temperature of the gas. If these two quantities are available, the emission can be immediately derived. It is also possible to merge two different Table Data structures to create a new one. Data in the resulting table can be treated as a single dataset. The merging capabilities and the mathematical operations give great flexibility in data analysis and representation.

Several built-in functions allow the user to perform a statistical analysis of a points distribution.

The **Scalar Distribution** function calculates the distribution of any quantity loaded in the Table Data and plots it as a histogram.

The **Correlation Filter** calculates the linear two-point correlation function of a point set. This is defined as the probability δP of finding a point in a randomly- chosen volume δV_1 and a point in another volume δV_2 separated by a distance r. The two-point Correlation Function of VisIVO is based on the 3D counterpart of the *Peebles & Hauser* estimator (Peebles 1974):

$$\xi_{PH} = \frac{DD(r)}{RR(r)}\left(\frac{N_{rd}}{N}\right) - 1, \qquad (1)$$

where N_{rd} is the number of points in an auxiliary random sample, $DD(r)$ is the number of all pairs of points with separation inside the interval $[r - dr/2, r + dr/2]$ and $RR(r)$ is the number of pairs between the data and the random sample with separation in the same interval. The random sample must have a density 2.5 times the density of the real dataset. The box is divided into a number N_{bin} of cubic subintervals. Then a frequency histogram of the pair distances of particles is constructed. The calculation of RR and DD is performed with a Monte Carlo integration.

The Fourier transform of the correlation function is represented by the **Power Spectrum**, which can be estimated by VisIVO as well. The power spectrum of a set of N massive particles can be calculated as

$$P(\mathbf{k}) = \langle | \rho(\mathbf{k}) |^2 \rangle. \qquad (2)$$

where $P(\mathbf{k})$ is the power spectrum, $\mathbf{k}$ is the three dimensional wave number: $k_i = 1/r_i$, with r_i indicating the i-th component of the spatial position of a point and $\rho(\mathbf{k})$ is the Fourier transform of the mass density field. The power spectrum provides the same statistical information of the correlation function, but it is much faster to compute. However, in the present implementation, periodic boundary conditions are required. Furthermore, in general, the spatial resolution is worse than that of the correlation function. In fact, in order to calculate the Power Spectrum, a Cloud in Cells algorithm distributes a constant value for each point (its mass) on a regular structured mesh with periodic boundary conditions. The mesh resolution sets the maximum wave number that the power spectrum can be calculated on. With this procedure, a mesh based mass density distribution is reconstructed and a fast FFT based approach can be used to estimate $P(\mathbf{k})$.

The last available analysis tool is for **Minkowski Functionals** (*MFs*). They describe the Geometry, the Curvature and the Topology of a point-set (Platzoder 1995). In a three-dimensional Euclidean space, these functionals have a direct geometric interpretation. The first Functional represents the volume V of a structure, the second one represents the surface area A and it is a measure of the geometry of the distribution. The third Functional corresponds to the integral mean curvature H of the structure's surface. It represents a measure of the distribution topology. The *MFs* algorithm implemented in VisIVO associates a *covering sphere* of radius r to each data point. The size, the shape and the connectivity of the spatial pattern, composed by the union-set of the spheres, change with the radius, which can be used as a diagnostic parameter. In particular, VisIVO computes the reduced values of the Minkowski Functionals, $\Phi(\mu)$ with $\mu = 0, 1, 2$, that are the ratio of the *MFs* of the actual distribution to the *MFs* of the same number of disjointed convex bodies. Their values always start from unity: for small radii, all the covering spheres are disjointed. In the 3rd functional $\Phi_3(r)$, the first zero provides an estimate of the percolation threshold. A spongy structure, like a Poisson distribution, gives lower values for $\Phi_3(r)$, while higher values indicate structures with few big filaments or tunnels.

2.3. VisIVO and VizieR

In the age of the Virtual Observatory, data collections are distributed between various sites. They are accessible to user applications via standard technologies such as the Web Service WSDL/SOAP protocol. One of the services exploiting this Web Service technology is VizieR, version 2 of which is available on the CDS servers. Although it is still in beta, the service will also be available at ADS, ADAC and CADC after the final release. VizieR is a database which archives, in an homogeneous way, thousands of astronomical catalogues gathered over decades by the CDS and participating institutes. The new web service interface gives access to the VizieR database of astronomical catalogues by adding four new methods to the old interface:

- coneCatalogs
- coneResults
- ADQLrequest
- getAvailability

VisIVO, using the Axis C/C++ library, implements an interface to the service. It is able to get the list of available servers using the getAvailability method, to get the list of valid parameters values to pass to the coneCatalogs and coneResults and using the last two methods to get metadata and data (in VOTable format) about catalogues depending on the given parameters.

In this way, VisIVO is able to query directly the VizieR web service to retrieve data from it and visualise them as if they were local data. The interaction with the service is transparent to the user. The user need only fill in specified fields with the parameters defining the data he wants to download. The result of this operation is a list of catalogues and, on selecting one of them, data can be visualised as if they were in a file or saved on the disk in the VOTable format.

3. Conclusions

Visualization cannot provide quantitative results, but it allows the user to have an immediate and intuitive approach to the data. The various 3D rendering techniques supported by VisIVO, together with the possibility of visualizing complementary quantities with colours, glyphs and vectors, allows the user to discriminate between data features at a glance, pointing out special characteristics and focusing on interesting regions. The software also implements a limited but effective set of statistical tools, that can be used to make quick estimates of the properties of a distribution. Mathematics functions let the user derive new fields starting from the original ones. VisIVO is being developed to follow the IVOA recommendations and standards, so that it is interoperable with the Virtual Observatory framework. Furthermore, it supports the PLASTIC protocol to allow the user to use the software together with other tools, such as Aladin or Topcat, which have complementary data analysis capacities to those of VisIVO. In this way the researcher can have a complete and customized cooperating set of tools that makes his/her research activity more and more efficient and focused on scientific issues.

New features of VisIVO will focus on exploiting new hardware architectures that are rapidly appearing in many desktop machines, such as 64 bit and multicore systems, subject to making effective use of such capabilities in the tool. The opportunity to use VisIVO in data centres and on dedicated visualization servers will drive the new releases of the code. Finally, VisIVO will be developed to integrate with the VO's Theoretical Data Archive framework. VisIVO will display a subset of the whole data file, which will generally be very large, and will allow the user to select a spherical or rectangular region and retrieve, through a remote service, the extracted sub-sample.

References

Bonnarel F. et al., 2000, A&AS, 143, 33

Ochsenbein F., Bauer P., Marcout J., 2000, A&AS 143, 221

Peebles P. J. E., Hauser M. G., 1974, ApJS, 28, 19

Platzoder M, and Buchert T. *Applications of Minkowski-functionals to the Statistical Analysis of Dark Matter Models.* 1995, A. Weiss, G. Raelt, W. Hillebrandt, and F. von Feilitzsch, Proc. of 1st SFB workshop on Astro-particle physics, Ringberg, Tegernsee , pages 251

Schroeder W., Martin K., Lorensen B., *The Visualization Toolkit An Object-Oriented Approach To 3D Graphics* 3rd Edition, 2004, Kitware Inc., ISBN-1-930934-12-2

Springel,V. 2000, MNRAS 1105, 364

Viceconti M. et al.. *The Multimod Application Framework* IEEE Proceedings of IV' 2004, 15

http://www.starlink.ac.uk/topcat/