

# Using CrystalVision

## Introduction

CrystalVision is an easy-to-use, self-contained Windows application designed as a platform for multivariate data visualization and exploration. It is intended to be a robust, intuitive, commercial-grade software. Key features include scatter plot matrix views, parallel coordinate views, rotating 3-D scatter plot views, density plots, multi-dimensional grand tours implemented in all views, stereoscopic capability, saturation brushing, and data editing tools as well as other capabilities that will be explained later in this guide. CrystalVision contains refinements of capabilities found in earlier software such as MacSpin, XGobi, Mason Hypergraphics, Explor4, and ExplorN. It has been used successfully with data sets as high as 20 dimensions and with as many as 500,000 observations. Data set size and dimension are limited in a practical sense by computer power and screen resolution. See Wegman (1995) for a discussion of these limits.

## Installing CrystalVision

CrystalVision is presented in two parts. The first part is a self-extracting file that will install CrystalVision on any contemporary MS Windows machine. It has been tested with Windows 95/98/NT/2000/XP. For installation with NT and 2000, the installer must be logged in as administrator or as a user with administrator privileges. In some early versions of 98 and in versions of 95 some additional .dll files may have to be installed. The second part is a zipped file of sample data sets.

The format for CrystalVision data sets is a text file. The first line contains

“variable:  $d$ ”

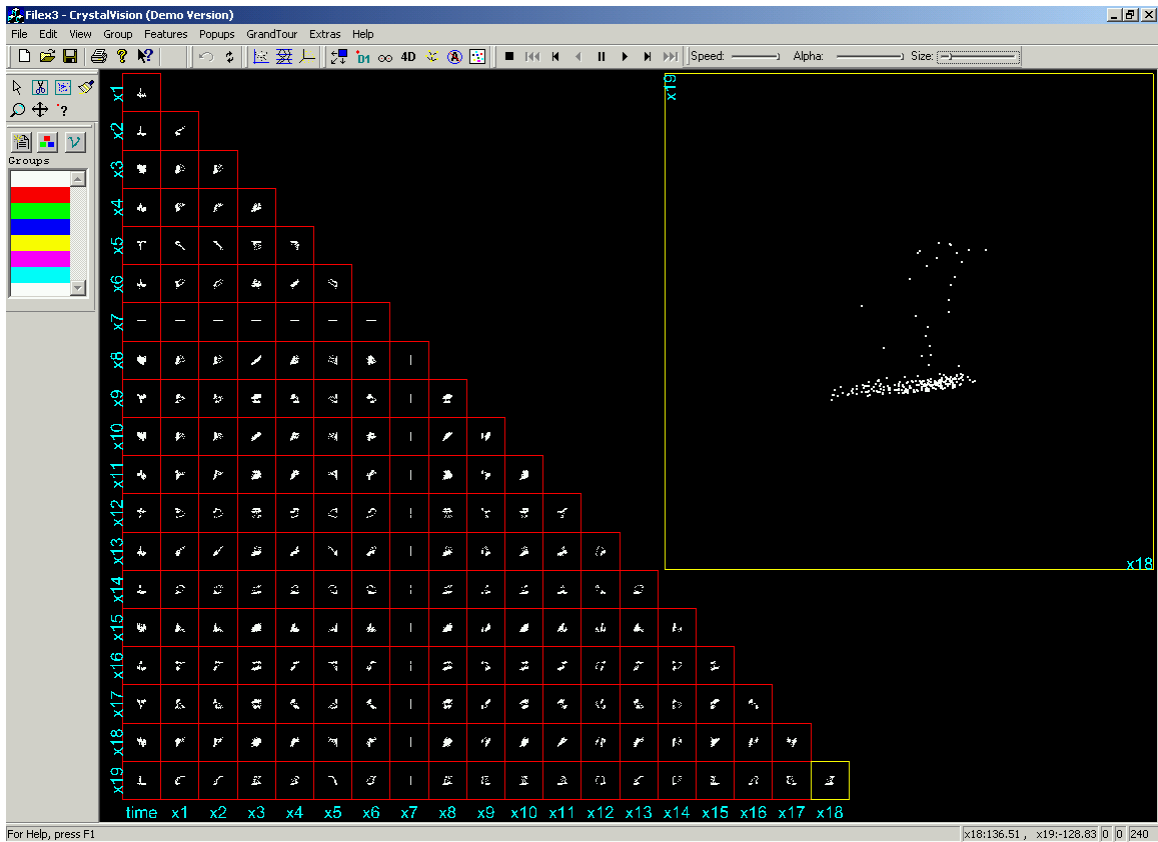
where  $d$  is the dimension of data set (number of attributes). The second line contains

“labels: ”

and the next  $d$  lines contain the labels for each of the variables (attributes). Both the word variable and the word labels must begin with lower case letters. Each of the remaining lines should contain the data, one line for each observation (case). There is no provision for missing data in this version of the software. Each observation may be either tab-delimited or space-delimited.

## Features

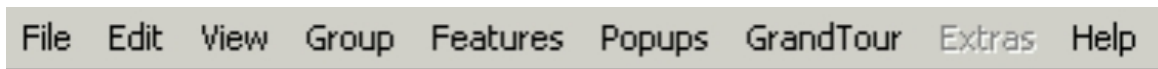
As mentioned earlier, CrystalVision has a number of capabilities that had been incorporated previously in special purpose software. There are drop-down and icon controls at the top and to the left of the main viewing screen. Figure 1 shows the full main screen as it appears.



**Figure 1. Opening Default Screen Shot of Crystal Vision**

The opening default view is the lower-left view of the scatter plot matrix. The upper right graphic is the highlighted scatter plot, which may be examined in more detail. Any small scatter plot in the scatter plot matrix can be highlighted and enlarged simply by mousing over the small scatter plot and left-clicking the mouse.

Individual functionality of the drop down menus is discussed below.



**Figure 2. The Drop Down Menu Items**

Left clicking on File will allow you to open a data file. Recently used data files are also listed. Also there is an Exit icon.

Left clicking on Edit allows you to copy the current graphic field to the clipboard. Once the image is copied to clipboard, it may be pasted into other applications such as MS Photo Editor, MS PowerPoint or MS Word. For MS Photo Editor, choose Paste as New Image and then save in an appropriate format. Usually a **.gif** file is an appropriate

compromise in terms of color resolution and file size. **Edit** also allows for some preferences.

Left clicking on **View** allows a choice of toolbar and status bar inclusion or exclusion and also allows a choice of stereoscopic algorithms and views. CrystalVision can exploit the stereoscopic visualization using the CrystalEyes shutter glasses technology with computers appropriately equipped. In particular, CrystalEyes can function as a virtual reality environment that uses the CrystalEyes shutter glasses. CrystalEyes drivers are available for Windows NT and Windows 2000 operating systems. See Wegman and Carr (1993) for a discussion of stereoscopic visualization.

Left clicking on **Group** allows you to import or export data with a group index. This is useful for recovering a partially done data analysis in which grouping had already been done.

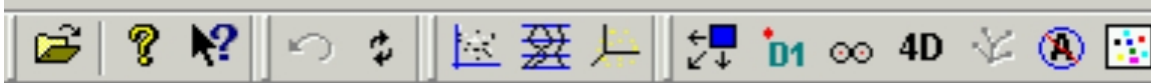
Left clicking on **Features** implements the visualization of a kernel density smoother of the data in two dimensions. In order for the density surface to be computed, at least two variables must be selected by the variable selection icon. Instructions for this are given below. The two-dimensional density surface is rendered in a 3-D box that may be spun by grabbing with the arrow icon and left clicking on the mouse.

Left clicking on the **Popups** icon allows the user to create additional graphics windows for scatter plot matrix views, parallel coordinate views, and 3-D scatter plot views. All of these windows may be used simultaneously and can be moved about as desired. If the grand tour is implemented, all windows will exhibit the same stage of grand tour.

Left clicking on the **GrandTour** icon will give current information on the Grand Tour status. The grand tour is implemented with a random geodesic algorithm. A recent discussion of grand tour algorithms may be found in Wegman and Solka (2002). The grand tour is implemented as a  $k$ -dimensional tour with  $3 \leq k \leq d$ , where  $d$  is the dimension of the data set. The variables upon which the tour is implemented may be chosen again by left clicking on the variables icon. See instructions below. The simultaneous multidimensional grand tour overcomes some of the handicaps of the original Asimov (1985) grand tour, which seemed to dwell for long periods in certain regions of the tour space.

The grand tour is an affine transformation (basically rotation) in high dimensional space. The columns of the rotation matrix give the linear combination of the variables that form the new variables. Rotated variables are denoted by adding an asterisk (\*) to the original variable name. If a variable is not involved in the grand tour, it will remain with no asterisk. In order to keep the rotated variables in the viewing window, a translation and scaling transformation is made. This information is also given when the **GrandTour** icon is left clicked.

Left clicking on **Help** displays this Guide.



**Figure 3. Part of the Tool Bar**



Left click on the **open file** icon will allow you to open a file. This can also be done from the drop down menu.



Left click on the **question mark** icon allows you to see these manual pages. It has the same functionality as the **Help** on the drop down menu.



Left click on the **undo** icon will return the screen image to its previous state.



Left click on the **refresh** icon will refresh the screen image.



Left click on the **scatter plot** icon displays the scatter plot matrix view. This is the default view when CrystalVision opens.



Left click on the **parallel coordinate** icon displays the parallel coordinate view.



Left click on the **3-D scatter plot** icon displays a 3-D scatter plot. In order for this plot to display properly, at least three variables must be selected under the variable selection icon. The 3-D scatter plot can be rotated, have grand tour with the 3-D box fixed or can have both the grand tour and rotation operating simultaneously.



Left click on the **enlarge** icon enlarges the scatter plot in the upper right corner of the scatter plot matrix view to full screen for more detail.



Left click on the **stereo eyes** icon implements the CrystalEyes stereo view for computers equipped with the appropriate hardware and software drivers. Stereoscopic display may be used for the 3-D scatter plot view or the 4-D view discussed below.



Left click on the **4D** icon will display the normal 3-D scatter plot with the addition of a tail added to each point. The angle of the tail represents a 4<sup>th</sup> variable. This was originally implemented in Dan Carr's Explor4 software. It is discussed in Wegman and Carr (1993). As with the 3-D scatter plot, the 4-D view can be rotated and/or have grand tour operational.



The no alpha icon relates to one of the key features of CrystalVision. CrystalVision features saturation brushing, which is a technique for dealing with larger data sets. Standard brushing brushes points (or lines) with a fully saturated color. In CrystalVision, we allow for brushing with a desaturated color, e.g. nearly black with only a small component of color. We use the  $\alpha$ -channel found in most modern graphics cards to add the small components of color. This creates a method for dealing with overplotting, typically a serious problem with large data sets. If a pixel is nearly black, there is little or no overplotting. If a pixel is fully saturated with one or more colors, there is considerable overplotting. Thus one can rapidly tell how much overplotting there is. Saturation brushing also gives a visual, hardware-based density estimate. Finally, because of  $\alpha$ -blending, we use additive color, e.g. red + green = yellow, red + blue = magenta, red + cyan = white, and so on. Thus clusters described by color can be distinguished when there is overplotting. See Wegman and Luo (1997) for the original discussion of saturation brushing and Wegman (2003) for a number of examples exploiting saturation brushing and the alpha channel. Use of the  $\alpha$ -channel hardware is the default. Left clicking on this icon removes the use of the  $\alpha$ -channel hardware and allows the program to perform like standard brushing. (Not recommended).



Left clicking on the reverse icon changes the background color from black to white. This has two uses. In journal publications, the publishers prefer to have a white background color rather than a black. Thus, this icon can be used to prepare graphics for publication. The default background color is black, which appears to be more desirable for studying data on a computer monitor. Contrast is higher and images are generally easier to study. Also when using saturation brushing, lightly overplotted items blend with the black background so that only regions of the graphic where there is moderate to heavy overplotting tend to show up. This visually discounts outliers. By reversing the background to white, the lightly overplotted pixels show up nearly black against a white background. Thus using this icon, one has the option of discounting or emphasizing lightly overplotted pixels (outliers).

The vertical bar icon is a standard windows icon that allows a tool bar to be moved. By left click and holding the mouse button, the tool bar can be dragged to a new location as desired. This is useful if CrystalVision is being used on a low resolution monitor.

NB: Icons not specifically mentioned are not implemented.



**Figure 4. Additional Tool Bar Icons**



This portion of the tool bar controls the grand tour and resembles standard controllers

for audio and video motion control. The grand tour is available in the scatter plot matrix view, in the parallel coordinates view, and in the 3-D and 4-D scatter plot views. It will operate simultaneously in all of the pop up views.



Left clicking on the **forward** icon initiates the grand tour. The tour will continue indefinitely. Step sizes are designed to be sufficiently small so that the tour conveys an impression of continuous motion. Very large data sets may cause the images to refresh slowly. This depends on the capability of the computer and the data set size.



Left clicking on the **pause** icon, you can stop the tour in order to examine a particular view in more detail.



Left clicking on the **one step backwards** or **one step forwards** icons will allow you to move individual steps in the grand tour either backwards or forwards. This is useful if you pass a particularly useful view and want to carefully recapture it.



Left clicking on the **original** icon will return the display to the original view before the grand tour was initiated. This is useful when using a BRUSH-TOUR strategy. See Wilhelm, Wegman, and Symanzik (1999) for a discussion of the BRUSH-TOUR strategy.



The **speed slider** icon controls the speed of the grand tour. The slider is operated by left clicking and holding on the slider tab and dragging it to the desired position.



The **alpha slider** icon controls the level of saturation. Fully to the right fully saturates all pixels, while fully to the left desaturates (makes black) all pixels. We tend to like a position somewhat to the left. Even when the slider is all the way to the right, the alpha blending is still on so that the additive color effects as described above still function. The slider is operated by left clicking and holding on the slider tab and dragging it to the desired position.



The **size slider** icon controls the pixel size. This slider works for the scatter plot matrix and the 3-D scatter plot views, but has no effect on the parallel coordinate view. The slider is operated by left clicking and holding on the slider tab and dragging it to the desired position.



The last set of icons on the left hand side of the display controls a number of other features of CrystalVision.

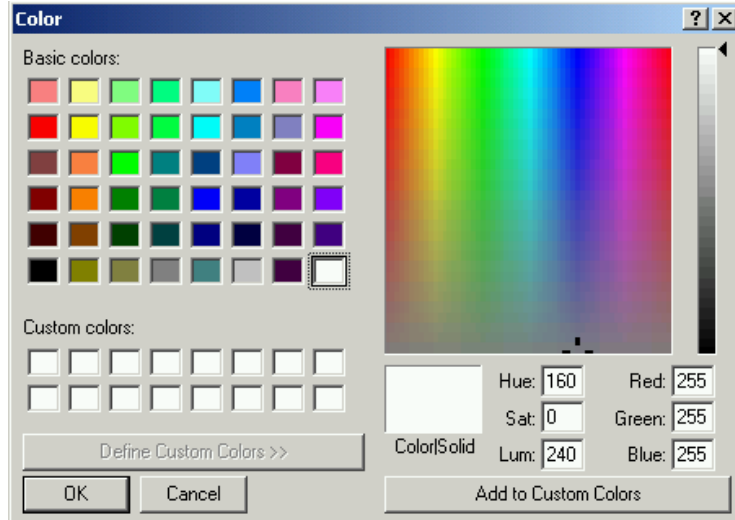
- The **arrow** icon is the default point icon and indicates the position of the mouse. The arrow icon also has a nice functionality in the parallel coordinate display. By left clicking on a variable name, holding, and dragging vertically, you can change the ordering of the parallel coordinate axes.
- The **scissor** icon is a cutting tool. When the **scissor** icon is engaged a rectangular box may be formed by pointing to a starting position, left clicking and holding and dragging to the final position. Anything that is inside the box will be cut from the screen image and deleted. This works in both the scatter plot view and the parallel coordinate view. In the parallel coordinate view, any line that goes inside the box will be cut.
- The scatter icon next to the scissor icon is a **cropping** tool. When the **cropping** icon is engaged, a rectangular box is formed just as with the scissor icon. The action is different however. With this icon, everything inside the box is preserved, everything fully outside is cut.
- The **brush** icon is used to brush the scatter plot or the parallel coordinate plot. The default color is white. To use the brushing tool, first choose a color by left clicking on a color, and then clicking on **brush** icon. A rectangular box is formed the same way as above. Everything touching inside the box will be brushed with that color.
- The **magnifying glass** tool can be used to magnify or make a graphics smaller and works with all of the standard views. To magnify, click on the **magnifying glass** icon, position the mouse pointer in the graphic area of choice and left click the mouse. Repeatedly left clicking will incrementally make the view larger. To make an image smaller, hold the control key down and left click the mouse. In either case, mouse pointer must be within the appropriate graphic.
- The **4-way arrow** icon is a centering icon. When this icon is enabled, you can left click and hold, drag and drop to recenter the image. This works in both the scatter plot view and the parallel coordinate view.
- The **question mark** icon is a query icon. When this is enabled, and the pointer is dragged over a point, its labels and values appear in a window just above the Windows tray at the bottom of the screen. This can be used to query outliers and other points.



The groups icon is used to create new grouping colors. There are seven original highly distinguishable colors: white, red, green, blue, yellow, magenta, and cyan. Red, green, and blue are the additive primaries and yellow, magenta, and cyan are the subtractive primaries. Red and cyan are complementary colors (they add to white). Similarly, green and magenta are complementary colors, and blue and yellow are complementary colors. Finally note that red + green + blue = white and also yellow + magenta + cyan = white. Thus these colors are extremely useful for categorical aspects of data analysis. Clicking on the groups icon adds additional brushing colors if the original 7 are not enough.



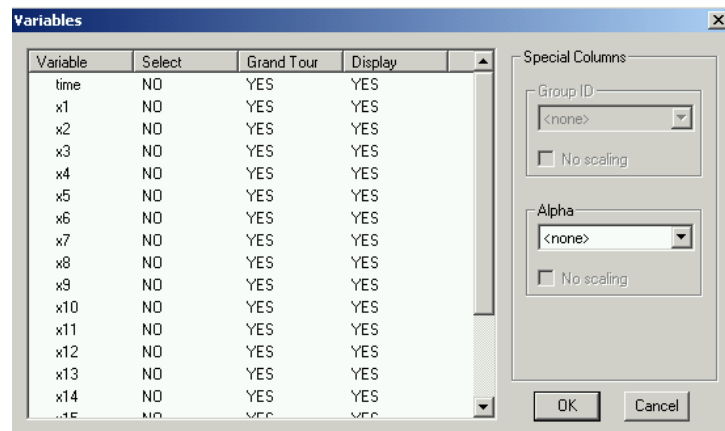
Left clicking on the color icon brings up the color chooser displayed here to the right. You can choose and/or modify the display colors. Note also that this chooser will display the RGB components of a chosen color as well as the Hue-Saturation-Luminance components. You can type in appropriate numbers in the boxes to determine particular colors.



Notice that each of R G and B can range from 0 to 255, i.e. 8 bits for each color yielding a total of  $2^{24}$  colors. Half-saturated colors are frequently useful, e.g. (red = 255) + (green = 128) + (blue = 255) = pink.



Left clicking on the variables icon brings up the variables selection panel displayed here to the right. The variables are listed along with three columns. The select column selects variables for inclusion in lower dimensional displays. Clicking on a NO turns it to YES and conversely. At least two variables must be



selected from the SELECT column for the density plot to be used. If more than two variables are selected, only the first two variables are used. At least three variables must be selected for the 3-D scatter plot to be used. Similarly at least four variables must be selected for the 4-D scatter plot to be used. In the GRAND TOUR column, multiple variables can be deselected. If so, they will not be used as grand tour variables resulting in a sub-dimensional tour. This is very useful if some variables are not appropriate for the

tour such as time in the panel illustrated here. This is also useful if one of the variables can be viewed as a response variable. Since the tour forms (orthogonal) linear combinations of the tour variables, the tour essentially effects a series of linear regressions of the response variable on the tour variables. The DISPLAY column allows you to remove a variable from consideration in the graphic. This could be because you want to do some visual dimension reduction or because the variable is defective or contaminated. A variable may be removed from the display, but not from the grand tour. The input box marked GROUP allows you to designate a data column as a grouping variable. The input box marked ALPHA allows you to select a data column as an alpha blending column. This is particularly useful if there is weighted data as for example arising from a stratified sampling procedure. The weights are reflected as the alpha values, which allow you to visually down-weight (i.e. desaturate) data points that have little weight from the sampling procedure.

Note in closing that Wegman (1990) discusses interpretations of parallel coordinates for exploratory data analysis, Wegman (1991) links the full-dimensional grand tour to parallel coordinate displays, and Wegman and Shen (1993) discusses fast pseudo-tour algorithm.

Wegman (2003), Wilhelm et al. (1999), and Wegman and Dorfman (2003) all contain examples of data analyses carried out with the tools described in this Guide.

## References

Asimov, D., (1985), "The grand tour: a tool for viewing multidimensional data," *SIAM Journal of Scientific and Statistical Computing*, 6, 128-143.

Wegman, E. J. (1990), "Hyperdimensional data analysis using parallel coordinates," *Journal of the American Statistical Association*, 85, 664-675.

Wegman, E. J. (1991), "The grand tour in k-dimensions," *Computing Science and Statistics: Proceedings of the 22nd Symposium on the Interface*, 127-136.

Wegman, E. J. (1995), "Huge data sets and the frontiers of computational feasibility," *Journal of Computational and Graphical Statistics*, 4(4), 281-295.

Wegman, E. J. (2003), "Visual data mining," to appear *Statistics in Medicine*.

Wegman, E. J. and Carr, D. B. (1993), "Statistical graphics and visualization," with D. B. Carr, in *Handbook of Statistics 9: Computational Statistics*, (Rao, C. R., ed.), Amsterdam: North Holland, 857-958.

Wegman, E. J. and Dorfman, A. (2003), "Visualizing Cereal World," to appear *Computational Statistics and Data Analysis*

Wegman, E. J. and Luo, Q. (1997), "High dimensional clustering using parallel coordinates and the grand tour," *Computing Science and Statistics*, 28, 352-360, 1997, republished in *Classification and Knowledge Organization*, (R. Klar and O. Opitz, eds.), Berlin: Springer-Verlag, 93-101, 1997. The paper with color images can be accessed with the following link: <http://www.galaxy.gmu.edu/papers/inter96.html>

Wegman, E. J. and Shen, J. (1993), "Three-dimensional Andrews plots and the grand tour," *Computing Science and Statistics*, 25, 284-288

Wegman, E. J. and Solka, J. L. (2002), "On some mathematics for visualizing high dimensional data," with Jeffrey L. Solka, *Sanhkyā (A)*, 64(2), 429-452.

Wilhelm, A. F. X., Wegman, E. J. and Symanzik, J. (1999), "Visual clustering and classification: The Oronsay particle size data set revisited," *Computational Statistics*, 14(1), 109-146. This paper can be accessed with the following link: <http://www.galaxy.gmu.edu/papers/oronsay.html>