July. 2016. Vol. 9. No. 01 ISSN2305-8269
International Journal of Engineering and Applied Sciences

www.eaas-journal.org

VISUALIZATION OF MULTIDIMENSIONAL DATA VOLUMES USING DIMENSIONAL REDUCTION TECHNIQUES

[1]MIGUEL DIOGENES MATRAKAS, [2]SERGIO SCHEER

[1]Student, Graduate Program of Numerical Methods in Engineering, Paraná Federal University UFPR
[2]Prof., Graduate Program of Numerical Methods in Engineering, Paraná Federal University (UFPR)
E-mail: [1]mdmatrakas@gmail.com, [2]sergioscheer@gmail.com

ABSTRACT

Description of a technique to display a multidimensional data volume such that the entire dataset is represented in the resulting image, not just a subset of the dimensions that are part of the data. The procedure involves implementation and testing, using General Purpose Graphics Processing Units, of dimensional reduction techniques to accomplish the projection of the multidimensional data in a color plane. The results indicate that it is possible to perform the visualization of a volume data by a size reduction technique for creating a color code for the data..

Keywords: ***Computer Vision, Scientific visualization, Dimensional Reduction, Multidimensional Scaling, Star Coordinates***

1. INTRODUCTION

The decision-making from a data set is influenced by the manner or method as their values are displayed. For sets with different relations between the variables visualization techniques with the appropriate characteristics should be used. In addition, data should be prepared, or pre processed, so that they can be represented correctly by the visualization process.

Real phenomena typically present multidimensional data, ie, have a large set of distinct features. Ideally, these projections count on the maximum possible dimensions or characteristics, to represent the original set, which is not always possible because graphical devices typically represent only two or three dimensions. Beyond that the understanding of a set with more than three dimensions is quite difficult 11 9.

In order for analysts to understand and draw conclusions from these data, it is necessary to created representations, or projections, in two or three dimensions, which can be correctly displayed on graphical devices.

This paper presents a study about some of the Dimensional Reduction (DR) algorithms that are used to create representations of *n*-dimensional sets with a small number of dimensions in order to facilitate its analysis and classification. In a second step, using DR techniques to process a data set representing a volume of *n* features, resulting in a new configuration, which simultaneously present the *n* characteristics, to be used to generate an image of the three-dimensional volume, represented in the data set. The aim of the study is to show the feasibility of representing all features present in the dataset in a single projection using GPGPU (General Purpose Graphics Processing Unit), providing a new tool for data visualization and analysis.

This article is organized as follows. In the next section we shall address Scientific Visualization and describe the characteristics of data volumes representation methods. In sequence are presented the concepts and descriptions of some DR methods. In the Dimensional Reduction section a method is presented for treating an *n*-dimensional data volume to generate it's graphical representation. Finally it is presented the final considerations on the proposed method, the results achieved and the next steps of the project.

2. SCIENTIFIC VISUALIZATION

Wright 11 defines Visualization as an interactive process to understand what led to, or produced, data, and not just a technical presentation of these data. To Ward and Grinstein 10 Visualization is the communication of information using graphical representations. These definitions are complementary, since they take into account the

origin of data, the presentation means and the interaction system that are employed.

For the visualization process, it is necessary to take into account that human beings with certain spatial reasoning naturally understands three dimensions. Consequently, understanding spaces with more dimensions, except for the special case of time is limited. Therefore, if it is necessary to represent more variables than can be accommodated with these restrictions, other resources should be used, such as colors, sounds, animation, or whatever else is available 11.

The set of operations to generate an representative image of a data volume consists of 4 10 :

- **Data Traversal** definition of the points in data volume, it provides the basis for the discretization of the visualization integral.
- **Interpolation** Normally the sampling points are different from the data grid, so it is necessary to reconstruct the continuous space from the grid to obtain the sample values.
- **Gradient Computation** The gradient of a scalar field is typically used to determine the local illumination.
- **Classification** Held usually by transfer functions is used to map properties of the data on optical characteristics, typically as a set of color values and opacity.
- **Shading and Illumination** Shading can be incorporated to the process by adding a term in the lighting visualization integral.
- **Compositing** It is the iterative process to determine the value of the visualization integral, which can be calculated either starting from the observer or in his direction.

Pao and Meng 8 address the problems of getting to understand a set of multidimensional and multivariate data, presenting as main tool the DR methods, allowing data to be viewed in 2D charts (projections), which allows analysts to understand more easily the relationships in the data. According to the authors there are three aspects in the understanding of multidimensional data:

- **Distribution of n-dimensional points** Knowing how the data points occupy the space by answering questions such as: Data distribution is uniform or in clusters? It follows the same distribution throughout the space, or is regular in a region and irregular in another?
- **Functional relationship** Whether there is a match between the values of the vector field of the input space and the space of the property values.
- **Categories creation** Clusters creation in the properties space. How the points in the data space relate to the categories? Elements close in the data space correspond to the same category in the properties space?

3. DIMENSIONAL REDUCTION

Dimensional Reduction is also part of the tools set available for data visualization and analysis, in which the number of dimensions exceeds the capacity of human understanding, or representation of a particular device. To illustrate a situation in which the use of DR is required, it is shown in Figure 1 an artificially generated three-dimensional block, comprising four scalar fields. The represented attributes have no specific physical significance, but were prepared to demonstrate the interaction between variables with different interpretation, scales and rates of change. The scalar shown in Figure 1a has a linear variation in only one axis. In Figure 1b, the representation shows that data exists only in one set of block faces and in Figure 1c the variation is linear as shown in Figure 1b, but in the other direction and with a larger amplitude. In Figure 1d it is seen that the scalar variation does not occur in only one direction, while maintaining a non-linear change rate. Figure 1e shows what may be considered as the solid volume and what is represented in Figure 1f refers to values lying in the region outside the solid.

In all represented dimensions, Figure 1a, 1b, 1c and 1d, it is possible to see that there are regions where there is no information to be represented, indicating that there are no values for the scalar being displayed. It is also important here to understand that the spatial coordinates that identify the values of scalar fields are also data set dimensions, so the example shown consists of 7 dimensions, the horizontal, vertical and depth axes, together with the scalars shown in Figure 1a, 1b, 1c and 1d (Figure 1e and 1f have been included to improve the understanding of the set, not making part of the actual data).

The DR procedure is to map the elements of a set with *n* dimensions for a representation that maintains in the best possible way, the relationships between elements and their groupings in a set with *m* dimensions, with m << n 2 9 1. Therefore, for a set of *h* elements

$X^n = \{x_i \in \mathbb{R}^n\}_{1 \le i \le h}$, an DR algorithm can be interpreted as a function defined as

$$\boldsymbol{f}: \mathbb{R}^{\boldsymbol{n}} \times \boldsymbol{T} \rightarrow \mathbb{R}^{\boldsymbol{m}} \quad \textbf{1}$$

that maps each x_i element into a new element y_i in space $\mathbb{R}^m$ 7.

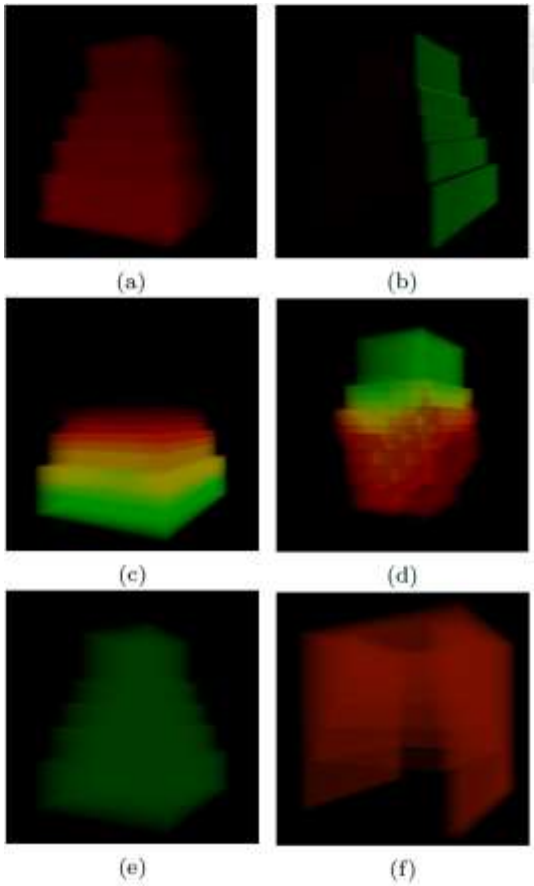

Figure 1 Volumes of Scalar Fields, Interior and Exterior Representing an Artificial Block Data Set

A multidimensional set after performing the size reduction should keep the neighborhood relations between the vectors, i.e., a set of points near the *n*-dimensional space must also form a set of neighbors in the projection data, that is, the *m*-dimensional space. Each of the DR methods presents a peculiarity regarding the disposal of the vicinity of the projected points 7.

There are dozens of methods used to make DR in the literature, classified according to the algorithm used to calculate the function *f*, presented in Equation 1. Some of the most used methods are the Classical Scaling or Multidimensional Scaling (MDS), Principal Components Analysis (PCA), Isomap, Maximum Variance Unfolding (MVU), Locally Linear Embedding (LLE), Stochastic Neighbor Embedding (SNE), Stochastic Proximity Embedding (SPE) and several configurations and models of Artificial Neural Networks. This list does not claim to be complete or qualify the methods, but present a set of the most cited, to expose the diversity of approaches to the DR problem.

Presented below are the description for the MDS method, along with a description of an iterative algorithm (SMACOF) that provides a solution to the MDS. The Star Coordinates system, a solution for data analysis that can be used for size reduction is also depicted.

3.1. Classical Scaling (MDS)

MDS, according to Borg2, consists of, from an array of elements *X* in the *n*-dimensional space, calculating a Δ^2 matrix containing the squares of these elements dissimilarities to then apply the operation called double centering consisting in calculating the matrix B_Δ given by

$$\boldsymbol{B_\Delta} = -\frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{J\Delta^2 J}$$

J being the centralization matrix given by: $J = I - n^{-1}U$, where *I* is the identity matrix, *U* is a matrix whose elements are equal to 1 and *n* is the number of dimensions of the original set.

Δ^2 matrix must be decomposed into its eigenvalues and eigenvectors, so that:

$$\boldsymbol{B_\Delta} = \boldsymbol{Q}\,\boldsymbol{\Lambda}\,\mathbf{Q}' = \left(\mathbf{Q}\,\boldsymbol{\Lambda}^{\mathbf{1}/\mathbf{2}}\right)\left(\mathbf{Q}\,\boldsymbol{\Lambda}^{\mathbf{1}/\mathbf{2}}\right)' = \mathbf{Y}\,\mathbf{Y}'$$

After the decomposition, it is considered the matrix formed by the first *m* eigenvalues greater than zero, called Λ_+ and Q_+ is a matrix formed by the first *m* columns of $\boldsymbol{Q}$, making the resulting array of coordinates to be:

$$Y = Q_+\Lambda_+^{1/2}.$$

This method minimizes the loss function given by

$$L(Y) = \|Y\,Y' - B_\Delta\|^2$$

3.2. SMACOF

As described in Borg and Groenen 2 it is an algorithm to minimize the stress function, whose acronym means "Scaling by Mojorizing a Complicated Function". This algorithm solves the Multidimensional Scaling by an iterative process. Therefore, from an array *X* of elements in the *n*-dimensional space and the Δ matrix

which is formed by the dissimilarity of these elements, the stress function represents the difference between the measures of the dissimilarities represented in the Δ matrix and the values of distance between the projections of the elements of X in *m*-dimensional space.

The stress function is written as:

$$\sigma_r(Y) = \sum_{i<j} w_{i\,j}\left(\delta_{i\,j} - d_{i\,j}(Y)\right)^2 \qquad \mathbf{2}$$

where w is a weight matrix, $\delta_{i\,j}$ are the elements of the dissimilarity matrix, $d_{i\,j}$ are the elements of the matrix formed by the distances between the elements of *Y*, which is the matrix formed by the projection of the *X* matrix.

The algorithm consists of, from an initial, nonrandom projection *Y*, calculate the difference between the distances using the Stress function, and while its value is greater than a precision limit, or a maximum of iterations is not reached, refresh the *Y* matrix using $Y^u = n^{-1}\,B(Y)Y$ Y if the weight matrix has all elements equal to 1, or $Y^u = V^+\,B(Y)Y$ otherwise. The matrix V^+ is the inverse of the weighted sums of the distances between the elements of *Y*, and $B(Y)$ is the matrix formed by the weighted ratio of the dissimilarity of the elements of *X* and *Y*.

The limitation of most MDS method applications is the need of $O(N^2)$ storage spaces and $O(N^2)$ operations, or computations, it can therefore be considered a problem limited by memory constraints 2.

3.3. Star Coordinates

The star coordinates aims to provide a system of easy understanding for the multi-dimensional data analysis by creating a 2D projection of the data set and providing a range of tools to enable users to explore the relationships in the input data set 6.

Kandogan 6 describes in his work the star coordinates as a plane in which the axes of the multi dimensional system are arranged in a circular fashion, sharing a common origin and separated from each other with the same angle. This system can be mapped to the Cartesian plane with the definition of a point of origin $O_n(x,y) = (o_x, o_y)$ and a sequence of vectors $A_n = \langle \vec{a}_1, \vec{a}_2, \cdots, \vec{a}_n \rangle$\$ that represents the axis of the *n*-dimensional space.

Each one of the D_j points from the input set is mapped to the Cartesian plane as the sum of the unit vector of each axis multiplied by the amount corresponding to this coordinate of the *n*-dimensional point, according to Equation 3.

$$P_i(x,y) = \left(o_x + \sum_{i=1}^{n} u_{xi}\left(d_{ji} - \min_i\right), \; o_y + \sum_{i=1}^{n} u_{yi}\left(d_{ji} - \min_i\right) \right) \qquad \mathbf{3}$$

where $D_j = (d_1, d_2, \cdots, d_n)$, $|\vec{u}| = \frac{|\vec{a_i}|}{\max_i - \min_i}$, $\min_i = \min\{d_{ji}, 0 \le j < |D|\}$ and $\max_i = \max\{d_{ji}, 0 \le j < |D|\}$.

4. VISUALIZATION OF A *n*-DIMENSIONAL DATA VOLUME

To attain the goal set in the first section, follows the description of the processing steps to be performed so that the dimensions of data sets, such as that visually presented in Figure 2, are displayed in a single image. In Figure 2 it is shown a simplified data volume, shown as a three dimensional array whose cells contain a vector of dimension 2 (two scalar values). In the case depicted in Figure 1 the matrix cells have a vector with dimension 4, represented respectively in Figure 1a, 1b, 1c and 1d.

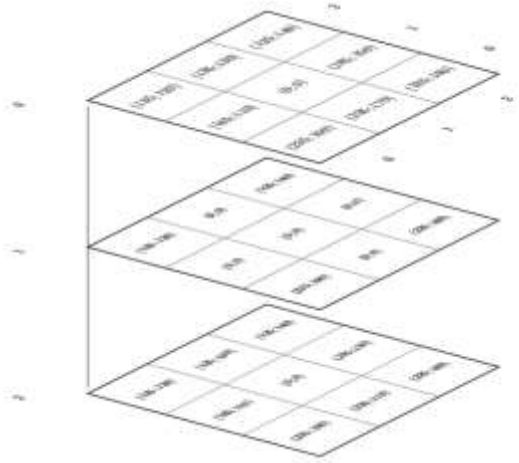

Figure 2 3D Representation of a Data Matrix Containing a 2 Dimensional Vector in Each Position

The first process step is to consider only the vectors, disregarding their position in the matrix which represents the spatial position of the data. To optimize the process, consider only one instance of each vector, i.e., when a vector is present in more than one cell of the matrix, it will count with only one instance on the dimensional

reduction step, to be performed in sequence. The result of this step applied in the Figure 1 matrix is shown in Equation 4.

$$\begin{Bmatrix}(235,100)\\(185,110)\\(135,120)\\(135,130)\\(135,140)\\(185,150)\\(235,160)\\(235,170)\end{Bmatrix} \quad \mathbf{4}$$

The data volume of the matrix depicted in Figure 1 has dimensions $[22; 22; 22]$, and a total of 982 different vectors. This matrix is provided to SMACOF algorithm without additional treatments to its values. The goal is to achieve a reduction to a two-dimensional space, where each coordinate is interpreted as an index into a color scale, thus forming a color plane. The projection of these 982 different vectors in 2D space using SMACOF is shown in Figure 3.

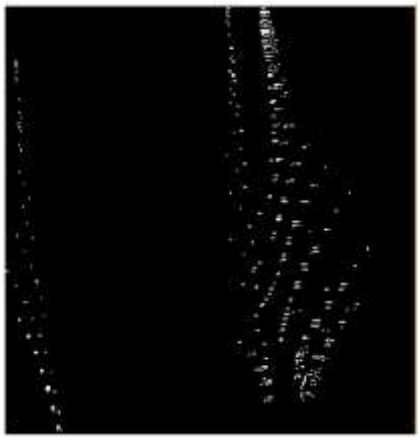

Figure 3 MDS Space Resulting From DR Using SMACOF Algorithm

In the proposed process, the only treatment performed on the data is a normalization, applied in the result of the DR achieved the SMACOF algorithm. The data is adjusted to the integers range $[0, 255]$, to facilitate it's interpretation on the color space, which is depicted in Figure 4. After running the DR, each element of the original data volume is replaced by the corresponding coordinates to its projection in the color plane, which will be used in the process of image generation corresponding to the content of all the data dimensions.

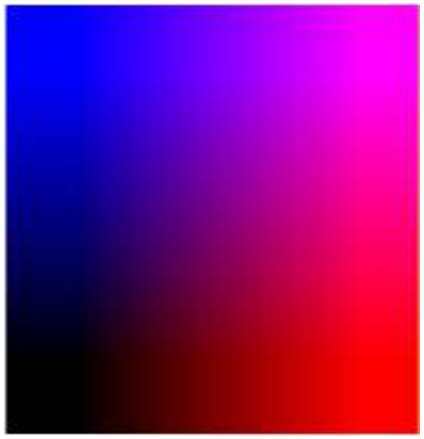

Figure 4 Color Plane Used to Represent SMACOF Algorithm Results

The result of this process applied to the data volume corresponding Figure 1 is depicted in Figure 5. The content of Figure 5a corresponds to the same view angle presented in Figure 1. In Figure 5b and Figure 5c are shown other viewing angles corresponding to the base and top of the block, respectively.

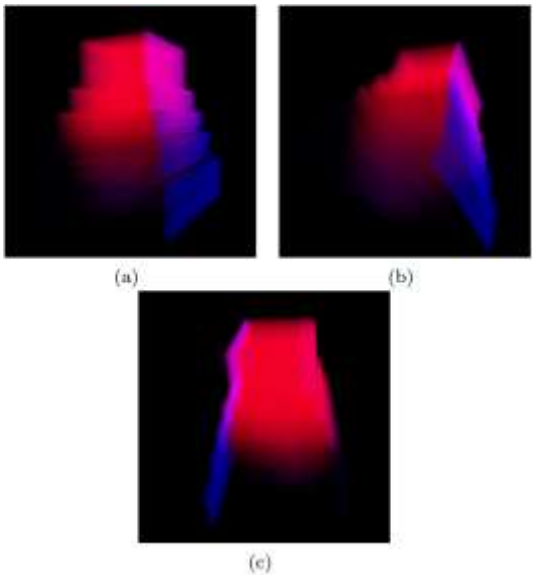

Figure 5 Rendering DR Results Using SMACOF in Data Shown in Figure 1

Applied the same procedure, but using star coordinates to realize the dimensional reduction, it is obtained a result with a different point distribution, according to the content of Figure 6. The data volume obtained, with the same viewing angles presented to the SMACOF algorithm, are shown in Figure 7. As by definition the origin of Star Coordinates space occupies the center of the projection plane, the color scale to be used need to take this feature

into consideration, with black occupying the center of the color plane, as represented in Figure 8.

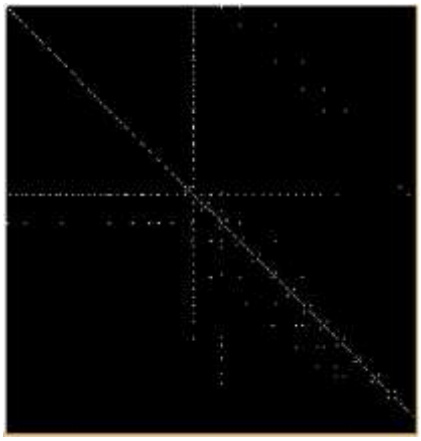

Figure 6 Space Resulting From DR Using Star Coordinates

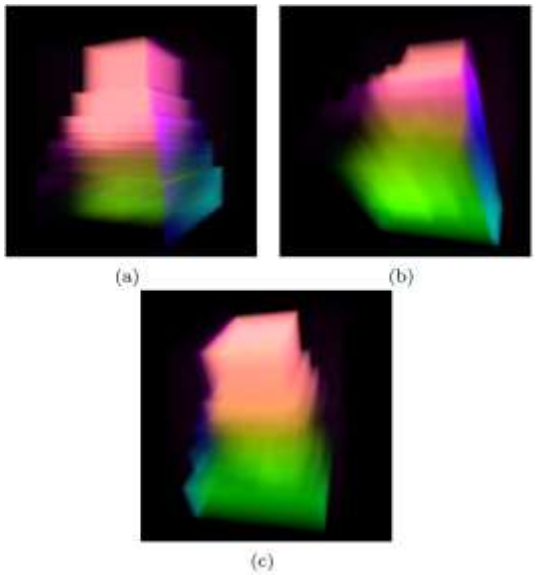

Figure 7 Rendering DR Results Using Star Coordinates in Data Shown in Figure 1

4.1. Hurricane Isabel data set

The methods were also tested with the Hurricane Isabel simulation data set, provided by NCAR (U.S. National Center for Atmospheric Research). This data set is composed of 48 time intervals, each corresponding to an array of $[500, 500, 100]$, the first dimension (X) corresponding to Longitude, the second dimension (Y) corresponding to Latitude and the third dimension (Z) is the elevation, or height from the sea level.

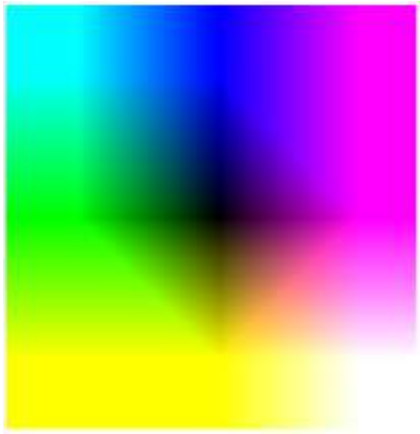

Figure 8 Color Plane Used to Represent Star Coordinates DR Results

In each one of the time intervals the following variables are available: cloud water, graupel, cloud ice, rain, snow, water vapor, total cloud (sum of cloud water and cloud ice), total precipitation (sum of graupel, rain and snow), pressure, temperature, wind velocity components in X, Y and Z.

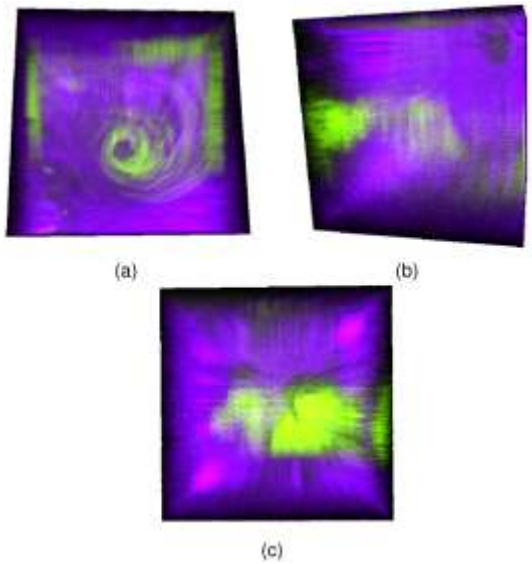

Figure 9 Rendering DR Results Using Star Coordinates in Hurricane Isabel Data Set

By applying the dimensional reduction and visualization process in this data set it is obtained images as shown in Figure 9. It is possible to see in the images some of the hurricane features as its vortex, cloud formations, in addition to the influence of the pressure profile and distribution of other variables. The content of Figure 9 is the result of RD considering values for pressure, cloud water, graupel, cloud ice, rain, snow, water

vapor and temperature. Only the thirtieth time frame was taken into account to generate the result depicted in Figure 9.

In order for the data matrix to fit on the graphics card available memory a $[27, 20, 30]$ size cutout of the data in the 30th time frame covering the region identified in Figure 10. This is necessary to accomplish the dimension reduction using SMACOF algorithm, due to the memory constraints imposed by the algorithm, as evidenced in Section 3.2. Only the amounts of graupel, ice, rain and snow were present in the matrix used to generate the result depicted in Figure 11.

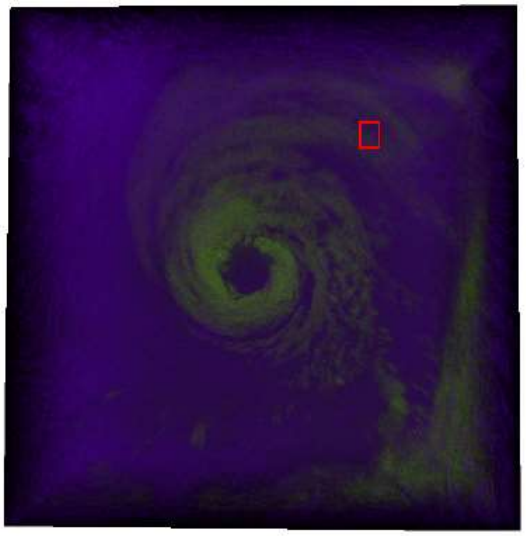

Figure 10 Cutout Region on Hurricane Isabel Data Set

In order to compare the results from both algorithms, the same matrix identified in Figure 10 was also processed using the Star Coordinates dimensional Reduction, resulting in the image presented in Figure 12.

The data volume of the matrix resulting from the a $[27, 20, 30]$ size cutout from the Hurricane Isabel data has a total of 13.198 different vectors. This matrix is provided to the SMACOF algorithm and to the Star Coordinates DR transformation to achieve the reduction to a two-dimensional space, interpreted as a color scale. The projection results obtained by the dimensional reduction using both techniques are shown in Figure 13. It is clear by analyzing the images produced (Figure 11 and Figure 12) and projections represented in Figure 13 that the DR algorithms focus on different relations of the data.

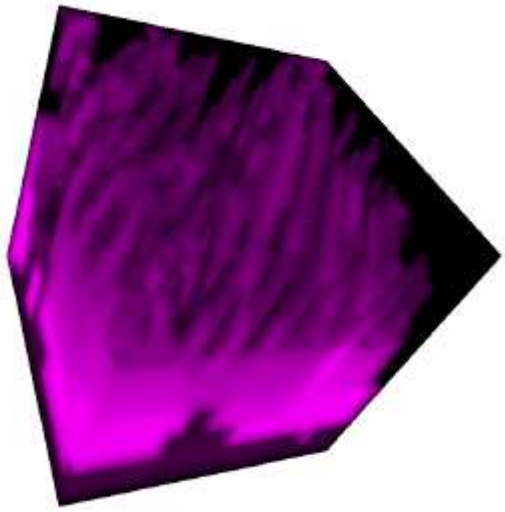

Figure 11 Rendering DR Results Using SMACOF in Region of Hurricane Isabel Data Set

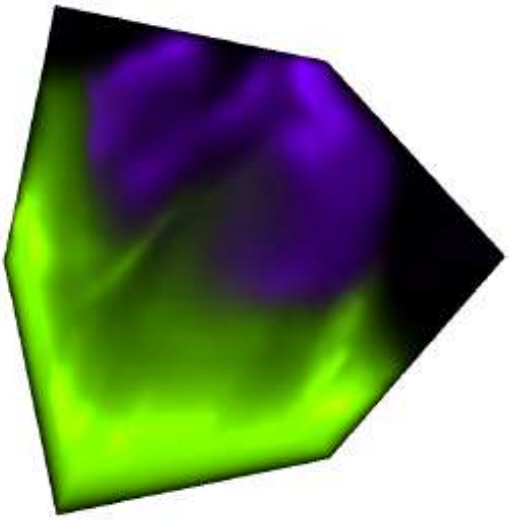

Figure 12 Rendering DR Results Using Star Coordinates in Region of Hurricane Isabel Data Set

Considering the algorithms execution on the artificial data block presented on Figure 1 the SMACOF algorithm took 69 calculation steps to reduce the initial stress value of 8,188,079,104 to 106,124,824 as the final stress value. Considering the cutout matrix of Hurricane Isabel simulation data the algorithm took 58 calculation steps to reduce the stress value from 258,132,590,592 to 1,392,438,144.

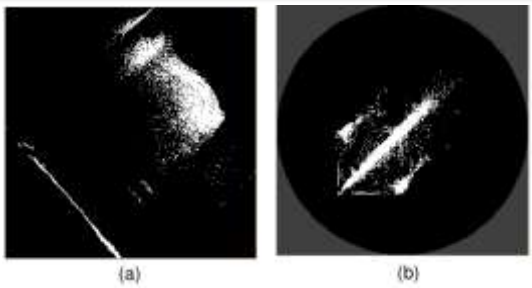

Figure 13 (a) Resulting Space Using SMACOF (b) Resulting Space Using Star Coordinates

The SMACOF runtime considering the artificial data block was 899 ms total processing time, or 13.03 ms step processing time. The same matrix was processed by the Star Coordinates DR in 5 ms. The processing time on the Hurricane Isabel cutout matrix with SMACOF was 73,636 ms with a step processing time of 1,248.07. The same matrix was processed by the Star Coordinates algorithm in 16 ms.

There is a difference in the total step processing time to the total that corresponds to the parts of the algorithm executed on the CPU, as the step processing time takes into account only the GPU processing time. This difference is linearly dependent on the matrix size.

5. FINAL CONSIDERATIONS

The present study shows that, using a GPGPU on a market graphics card, it is feasible to carry out the display of an *n*-dimensional data set showing all its features simultaneously, since the dimensions are reduced, so that in each volume position being represented are the coordinates of a color code to be used in the output image generation process.

It is not in the scope of this work to interpret the results regarding the interaction of the different dimensions in the resulting image. So no other comments regarding the contents of each resulting image will be made.

The result obtained, according to considerations regarding what is illustrated in Figure 3, Figure 6 and Figure 13, demonstrates the feasibility of using a size reduction technique in order to achieve a representation of a set of multidimensional data, enabling the generation of an image containing a representation of all the set information.

The results obtained with both the MDS (SMACOF algorithm) and the Star Coordinates, show distinct regions of each of the original dimensions in the resulting image, considering both data sets used.

Analyzing the artificial data block, the distribution outcome of the MDS projections reduces the participation of the corresponding content to Figure 1c. However, the images generated from the DR performed with the Star Coordinates, it is possible to clearly see the influence of the four dimensions shown in Figure 1.

On the contrary, in the case of the cutout matrix from the Hurricane Isabel simulation data set, the results obtained with the SMACOF algorithm show more details then the one resulting from the Star Coordinates dimensional reduction.

As each of the DR algorithms have different objectives, the results reflect these characteristics. This highlights the fact that in a system for analysis of three-dimensional data, as proposed in this work, a wider range of DR techniques should be available, so the analysts can choose to use the one that best suits the input data and also the analysis type to be carried out.

Confirmed the feasibility of the process, the next steps in the project development involve tests with other Dimensional Reduction techniques and different data sets, so a categorization on the characteristics of both can be obtained.

Also should be implemented tools for manipulating data sets and results, such as changing the number of dimensions to be used in the DR stage, determination of cutting planes in the volumes, to facilitate inspection of the inner regions and the determination of value ranges to be considered on the display. Tools to manipulate projected points in the color plane, such as selection, rotation and classification, will be important additions to aid data analysis too.

6. ACKNOWLEDGMENTS

The authors will like to thank the support given by the Itaipu Binational (the hydroelectric power plant and dam company), to the Graduate Program in Numerical Methods for Engineering (PPGMNE - *Programa de Pós Graduação em Métodos Numéricos em Engenharia*) at UFPR , as well as to thank the Center for Advanced Studies on Dam Safety (CEASB - *Centro de Estudos Avançados em Segurança de Barragens*) of the Itaipu Technological Park (PTI - *Parque Tecnológico Itaipu*).

The authors also will like to thank Bill Kuo, Wei Wang, Cindy Bruyere, Tim Scheitlin, and Don Middleton of the U.S. National Center for Atmospheric Research (NCAR), and the U.S. National Science Foundation (NSF) for providing the Weather Research and Forecasting (WRF) Model simulation data of Hurricane Isabel.

REFRENCES: PROPER APA OR MLA STYLE

1. L. Adhianto, S. Banerjee, M. Fagan, M. Krentel, G. Marin, J. Mellor-Crummey, N.R. Tallent, "HPCTOOLKIT: Tools for performance analysis of optimized parallel programs", Computation Practice and Experience, 2013, pp. 662-682.
2. S. Bae, J Qiu, G Fox, "Adaptative Interpolation of Multidimensional Scaling", Procedia Computer Science, jan. 2012, volume 9, pp. 393-402, DOI 10.106/J. PROCS.2012.04.042.
3. Borg, P. J. F. Groenen, "Modern Multidimensional Scaling", 2005, Springer.
4. K. Engel, M. Hadwiger, J. M. Kniss, A. E. Lefohn, C. R. Salama, D. Weiskopf, "Real-time volume graphics", 2006, A K Peters Ltd.
5. Hurricane Isabel WRF Model Data, 2009, available at <http://www.vets.ucar.edu/vg/isabeldata/readme.html>.
6. E. Kandogan, "Star Coordinates: A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions", 2000, IEEE Information Visualization Symposium.
7. R. M. Martins, D. B. Coimbra, R. Minghim, A. C. Telea, "Visual analysis of dimensionality reduction quality for parameterized projections", 2014, Computers and Graphics (Pergamon), pp. 26-42.
8. Y. Pao, Z. Meng, "Visualization and the understanding of multidimensional data", Engineering Applications of Artificial Intelligence, 1998, pp. 659-667.
9. L. J. P. Van Der Maaten, E. O. Postma, H. J. Van Den Herik, "Dimensionality Reduction: a Comparative Review", Journal of Machine Learning Research, 2009, pp. 1-41.
10. M. O. Ward, G. Grinstein, D. Keim, "Interactive Data Visualization", 2015, 2nd edn., CRC Press.
11. H. Wright, "Introduction to Scientific Visualization", 2007, Springer, Hull - UK.