Entropy Based Information Visualization of Scientific Data

Guoqing Wu, Yi Cao and Junping Yin

Institute of Applied Physics and Computational Mathematics, Bei Jing, China

***Abstract*—Visualization of large scale time-varying scientific data has been a challenging problem due to their ever-increasing size. Identifying and presenting the most informative (or important) aspects of the data plays an important role in facilitating an efficient visualization. In this paper, an information assisted method is presented to locate temporal and spatial data containing salient physical features and accordingly accelerate the visualization process. Based on information measures, the method adaptively picks up important time steps and sub-regions with the maximum information content so that the time-varying data can be effectively visualized in limited time or using limited resources without loss of potential useful physical features. The experiments on the data of radiation diffusion dynamics and plasma physics simulation demonstrate the effectiveness of the proposed method. The method can remarkably improve the way in which scientists analyze and understand large scale time-varying scientific data.**

***Index Terms*—About four Time-varying data, scientific data, information visualization.**

I. INTRODUCTION

With growing capability of supercomputers, scientists are able to simulate sophisticated physical phenomena with ever-increasing scale and accuracy. At the same time, these numeric simulations also lead to the generation of extremely large amount of multi-dimension time-varying simulation data. For example, a typical interacting 3D simulation of laser incident into plasma (see Fig. 1) using a high performance parallel processing computer may includes thousands of time steps, each time step representing a regular grid data set at a resolution of 1000^3. Hence, the total size of resulting data may reach TB scale. The size and complexity of the data in these scientific domains is such that it is impractical to store and visualize the full extent of the simulation data all at once. The situation is becoming worse, making it difficult or even impossible for scientists to effectively explore and understand simulation data. As a result, useful information is often overlooked. Therefore, it's necessary to extract data containing interested or important features from these massive scientific time-varying data and allow the scientists to visualize the most salient information in their simulation result without skinning the entire data.

Manuscript Received on May 12, 2011
E-mail: lindazsj@163.com

Time-varying scientific data are dynamic in nature, and thus visualization is a challenging task and demands novel thinking and new techniques. In fact, scientists are not interested in a stationary process but usually focus on time points or regions when or where abnormal changes take place in a non-stationary process. Hence, one solution is to extract the data that contains rich temporal and spatial features and visualized them. In this way, the most informative aspects of the original massive data are revealed while trivial parts are deemphasized, thus scientists can purposefully focus their attentions on the dynamic features and analyze the scientific data more effectively. A viable method for extracting such important features is to utilize intrinsic information quantity contained in the data. The major benefit is that by using information-theoretical measures, the salient feature can be located without explicit feature description, and allow more efficient implementation.

In this paper, we present an information assisted visualization approach to locate and visualize the most informative time steps or sub-regions from time-varying scientific simulation data using information-theoretic measures. This is achieved by utilizing several measures from information theory, including Kullback-Leiber distance and off-line marginal utility to detect the time steps that the underline phenomena change most, and normalized entropy to locate where the features exist in a certain time step. Most important time steps are the ones that have most information dissimilarity and provide surprising new information. Similarly the idea hold for the selection of sub-regions. For a certain time step, most important sub-regions are the ones that have more information quantity.

Our method automatically identifies time-space anomalies and adaptively visualizes selected important time steps and sub-regions that contain physical features while discarding non-important ones. Meanwhile, the method reduces the data size and accordingly cuts down rendering time cost. Compared with the common practice of importance-driven visualization, our method extracts time steps and sub-regions capturing the maximal amount of information. We use applications from physics to demonstrate the effectiveness of our method.

The remainder of the paper is organized as follows. In Section 2, we review previous related work. In Section 3, we introduce some basic concepts of the information theory used in our work. The information assisted visualization method for time-varying scientific data is then presented in detail in Section 4. We give some experiment results in Section 5 to validate the effectiveness of our method. In Section 6, the

potential benefits of our techniques are discussed.

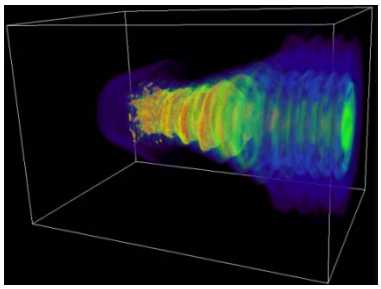


Fig. 1. Interacting simulation of laser incident into plasma

II. RELATED WORK

Time-varying data visualization has always been an important research topic. Shen [18] developed time-space partitioning tree to capture spatial and temporal feature from time-varying data. Fout [8] presented a solution of data reduction in supercomputing environments where simulation time-varying data occupies tremendous amounts of storage. Wang [22] have experimented with application-driven approach to compress large scale time-varying volume data. Shen [17] presented several algorithms for visualizing large scale time-varying scientific data including: (1) Lossless spatial-temporal data encoding and indexing schemes; (2) Coherence based accelerated visualization algorithms; (3) Time-varying feature enhancement and tracking algorithms. Their goal is to minimize data transfer and visualization cost, and to detect and track important time-varying feature.

There is an extensive literature in computer graphics and animation to address the objectives of identifying important features in time-varying data. Wang [21] evaluated the importance of data around a spatial local neighborhood (i.e., a data block) in the joint feature-temporal space. Using the conditional entropy, they derived importance curve for each data block which characterizes the local temporal behavior of the block. Based on different temporal trends exhibited by importance curves and their clustering results, they demonstrated effective visualization techniques to visualize and understand dynamic temporal features of time-varying data. Viola [20] proposed an importance-driven solution for automatic focusing on features within a volumetric data set. They discussed the necessary pre-processing steps before a visual inspection, including localization of the most expressive viewpoints. Caban [5] introduced a texture-based feature tracking technique capable of tracking multiple features over time by analyzing local textual properties and finding correspondent properties in subsequent volumes.

Information theory has been widely used in visualization and computer graphics. Wang and Ma [21] summarized a few directions of information and knowledge assisted analysis and visualization of large scale data. They consider that the future of visualization lies in the development of information and knowledge driven solutions to handle with ever increasing mounts of data. Chen [6] reported theoretic findings on whether information theory can become one of the theoretic frameworks of visualization. They also outlined the broad correlation between visualization and the major applications of information theory. Xu [25] presented an information-theoretic framework for flow visualization. With their framework, the distribution of streamlines is controlled by the information content of the data so that more streamlines will be seeded in regions with higher information content. Pescape [15] proposed an off-line Entropy-based methodology for reducing large network traffic data sets obtained by measures over real networks. Their approach is complementary to general sampling approach and more useful when large data sets are used to characterize network traffic without losing sensible information. Song [19] and Dasu [7] presented a general information-theoretic approach for non-parameter change detection, which works for multidimensional data streams. The approach use KL-distance to generalize traditional distance measures in statistics and works efficiently and accurately.

It is also worth mentioning that information theory has been widely used in good view points selection [2, 9, 10] and region of interest [2, 4, 24]. Bordoloi [2] studied the problem of finding a good static viewpoint for time-varying data set using Shannon entropy. Ji [10] improved the method in [2] and used static view selection method and a dynamic programming approach to select both static and dynamic viewpoint for time-varying data. Feixas [9] integrated Shannon entropy and KL-distance and proposed an efficient algorithm for viewpoint selection and mesh visibility and saliency. In [1, 4, 24], the authors proposed techniques based on information-theoretic measures to locate regions of interest in images or videos.

The goal of our work shares partly similarity with [21] (such as time step selection) and [5] (such as feature tracking). Unlike [21], we locate feature's time step in the view of change and surprising information. Our solution is simpler and more general than [5] in locating the feature's region. Caban [5] defined textural metrics to search different texture pattern while we utilize intrinsic information content instead of explicit feature description.

III. INFORMATION-THEORETIC MEASURE

Comparing with common statistics, such as mean and variance, information-theoretic measures capture more general aspects of data and thus have advantage of robust and generality. More important, information-theoretic measures are defined independent of the inherent dimensionality of the data and even the spatial nature of the data. Thus, many problems are being viewed through the lens of information theory. In the following, we review some basic concepts of information theory [16] used in our work.

3.1 Definition

Information theory provides a complete theoretical framework to quantitatively measure the information content from a distribution of data values. Shannon entropy is a fundamental measure in information theory. Given a random variable X with a sequence of possible outcomes x ,

$x \in A = \{x_1, ..., x_n\}$. If we know that the probability for random variable X to have the outcome x_i is $p(x_i)$. $p(x)$ is referred to as a probability mass function. Then the information content for the random variable can be computed using Shannon's entropy as

$$H(X) = -\sum_{x_i \in A} p(x_i) \log p(x_i) \quad (1)$$

Where A is the space composed of all possible outcome x_i of the random variable X. The quantity $-\log(p(x_i))$ is defined as the information content which is associated with outcome of x_i . *H(x)* is the average information content of all x_i . The unit of information is bits when logarithm taken base 2. Shannon's entropy is to measure the uncertainty of a random variable. An important property of the entropy is that *H(x)* is convex and reaches its maximum when $p(x_i)$ is equal for all X_i

A natural way for detecting changes between different time steps is to model multi-dimensional data via distribution and use some distance characteristics. A measure that is one of the most general ways of representing this distance is the relative entropy (also called Kullback-Leibler distance). The relative entropy between two probability distribution *P* and *Q* is defined as [13]

$$KL(p \,||\, q) = \sum_{x_i \in A} p(x_i) \log \frac{p(x_i)}{q(x_i)} \quad (2)$$

The KL-distance has many properties that make it ideal for estimating the distance between distributions: (a) The KL-distance generalizes standard tests of difference like the t-test, chi-square [7]. (b) It is an example of an α -divergence and have various geometric invariance properties [13]. (c) The KL-distance measures information dissimilarity. Therefore, compared with Euclidean distance in R^n , it allows us not only to measure the difference between time steps, but also attribute a meaning to this value.

Besides, the KL-distance was also exploited another meaning in [3] to measure the marginal utility of adding a new data set to an aggregate current data sets. The concept of marginal utility comes from economics. For an intuitive explanation, one can see that the first cup of water quenches a man's thirst best while the second or third cup of water has less and less thirst-quenching effect. The more cups of water, the less effectiveness of quenching his thirst. Similarly, for a time-varying scientific data, all time steps may not be equally important for scientist to understand physics process. It is also in accord with the rule of marginal utility descending. The marginal utility of the time step S^m can be estimated by considering the reduction in uncertainty (i.e. new additional information) provided by it. The reduction in uncertainty for each outcome x_i , after the time step S^m , is

$$-\log(p(x_i^{m-1})) + \log(p(x_i^m)) = \log(\frac{p(x_i^m)}{p(x_i^{m-1})}) \quad (3)$$

Where $p(x_i^j)$ is the probability associated with the outcome x_i after the conclusion of time steps $S^1, S^2, ..., S^j$. The marginal utility of the time step S^m is then defined as the mean reduction in uncertainty caused by the addition of the time step to the current time steps. In [3], an off-line marginal utility is presented. The off-line marginal utility of the time step S^m , with $m \le n$, is defined as

$$U^n(S^m) = \sum_{x_i \in A} p(x_i^n) \log \frac{p(x_i^n)}{p(x_i^m)} \quad (4)$$

It quantifies how much additional new information S^m provides relative to all time steps. n is the total number of time steps. The off-line marginal utility considers the marginal utility of current time step from the perspective of the whole time-varying data. Clearly, the utility of a supplemental time step decreases as it does not bring new information.

3.2 Entropy of Data Set

The information theory is associated with probability distribution. Evaluating how much information content a data set contains is equivalent to examining the abnormity of the corresponding data distribution. A distribution with sharp peak has a low entropy value, while a dispersed distribution yields a high entropy value. For an extreme example, visualization of a sub-region with a single value is a monotone color picture and offers nothing information to an observer.

To measure the information content of a data set S whose elements are float, we need to construct probability distributions from data to approximate the probability mass function. Let us consider a data set $\{d_i\}_{i=1}^N$, where N is the length of the data set. If d_i is a continuous variable: (i) we consider a finite interval (a,b) such that $a = \min\{d_i\}$ and $b = \max\{d_i\}$; (ii) we divide the interval (a,b) into n nonintersecting subintervals $\{l_i, l_{i+1}\}_{i=1}^n$ of equal length $L = (b-a)/n$; (iii) then the probability $p(x_j)$ of the subinterval (l_i, l_{i+1}) is given by

$$p(x_j) = \frac{\#\{d_i \mid d_i \in (l_j, l_{j+1})\}}{N}, i = 1, ..., N, j = 1, ..., n$$

where the symbol # indicates the number of elements in the set.. It is noteworthy that probability distributions, and hence the information-theoretic measures are affected by the number of histogram bins. An experiential formula of n is $1.87*(N-1)^{0.4}$ [11].

IV. INFORMATION ASSISTED VISUALIZATION

4.1 Information Flow Model of Time-varying Data

Before introducing the algorithm of selecting informative or

important time steps, we firstly construct an information flow model for time-varying scientific data as the foundation of our work. In the time-varying data, each time step provides a certain quantity of information about scientific phenomena. More time steps, more information we gained. As the time passes, the entire time-varying data gradually provides all information about how phenomena evolve. The information provided by sequential time steps just likes an information flow. Let us consider a sequence of n time steps $S^1, S^2, ..., S^n$. Φt is the information provided by the conclusion of the time steps $S^1, S^2, ..., S^t, t \le n$. Then Φt is a field with properties: (1) Φt $\subset$ Φt+1. (2) If $A \in$ Φt, and $B \in$ Φt, then $A \cup B \in$ Φt , $A \cap B \in$ Φt , $A \setminus B \in$ Φt . Define a filtration $*$ which is the collection of fields $*$= {Φ1, Φ2 …, Φt…, Φn}, then $*$ can be used to model the flow of information in time-varying data. With the passage of time, an observer knows more and more detailed information about simulated physical phenomena, that is, finer and finer partition of true scientific discipline, and meanwhile the quantity of new information tends to zero gradually. Corresponding to union, intersection and difference operations of field (i.e. $\cup, \cap, \setminus$), we can utilize the operations $H(\cdot,\cdot), I(\cdot,\cdot), H(\cdot|\cdot)$ to examine the information flow between different time steps.

4.2 Information Assisted Selection of Time Steps

Visualization of massive scientific data takes a long time to load, transform and render the data. Therefore, it is helpful to select a fraction of time steps from a long time sequence to decrease the computational resource demands. In general, we can select and animate 1 out of every 10-200 time steps according to available computing resources and time budget we can afford. The followed problem is how to select these time steps effectively and reasonably. Ideally, scientists expect to obtain the maximum amount of information from the least number of time steps. The most common way of time step selection is to select time steps uniformly from the time step sequence (e.g., select one every k time steps). The uniform selection method, although convenient, may lose important time steps since physical phenomena sometimes evolve quite unevenly along the time line. A more rational way is to select time steps according to the quantity of important physical features.

As time tends to infinity, scientific phenomena tend to a stable state, and new time steps provide less and less new information. When dealing with an additional time step, we want to compute how much new information we gained from it and then decide whether to extend current animation with it. As mentioned in Section 3, the off-line marginal utility of a time step clearly indicates how much new information it contributes. When the off-line marginal utility of a new time step is less than a given tolerance ε (e.g. 10^{-2} in our experiments), it is unnecessary to visualize the followed time steps.

When selecting time steps from a time-varying data, we also need to take information similarity into account, especially when scientific phenomena change slowly over time. We calculate KL-distance of neighbor time steps to check whether its information is remarkably different from its neighbors. Exploiting information similarity between consecutive time steps can avoid redundant visualization.

Next, we propose an information-theoretic algorithm for effectively selecting key time steps that contain important physical features. The first time step is always selected. Then, we calculate off-line marginal utility of each time step to check when to stop visualization. Hereafter, we partition truncated time steps into k-1 segments while KL-distances of each segment are nearly equal. k is the number of selected time steps and chosen by user according to rendering time budget. Finally, select one time step from each segment. The algorithm is summarized as follows.

Algorithm 1. Select most informative time steps
// input: n consecutive time steps $S^1, S^2, ..., S^n$;
k is the number of selected time steps.
// output: indexqueue of selected time steps;

push 1 to indexqueue;
for (i = 1,...,n-1) {
compute off-line marginal utility $U^n(S^i)$ of ith time step;
if ($U^n(S^i) < \varepsilon$) {
flag = i;
break;
}
}
for (i = 2,…, flag)
compute $KL(S^{i-1} \| S^i)$;
$KL_{average} = \frac{1}{k-1} \sum_{i=2}^{flag} KL(S^{i-1} \| S^i)$;
count=1;
for (i = 2,…, flag) {
if $(\sum_{j=2}^{i} KL(S^{i-1} \| S^i) > (count * KL_{average}))$ {
push i to indexqueue;
count++;
}
}
return indexqueue;

The output of Algorithm 1 is a list of un-uniform time steps, which is a temporal coarse version of the time-varying data and easier to be visualized. Animating the time steps selected by Algorithm 1 presents a visual summary of the original time-varying data. When users input a proper parameter k, a number of key time steps are selected, and the resulting animation conveys the informative aspects of the original data. In addition, the selected time steps have less information redundancy and can be regarded as a representative of the original time-varying data. Therefore, such selected time steps enable scientists to analyze the data more effectively or browse the entire time-varying data more efficiently.

4.3 Information Assisted Location of Sub-regions

Another interesting problem is to identify or locate

sub-regions with important physical features in a certain time step. Given a limited time budget for rendering a data set, we can allocate more time to render important sub-regions in higher resolution. When visualization resources are limited, we can also render important sub-regions with higher priority and abandon unimportant ones.

To locate sub-regions with important physical features, we take a block-wise approach and partition the volume data into spatial blocks. For example, we first partition the data set of the tth time step into m rectangle sub-regions (see Fig.2). The block size may affect the effectiveness of our method. Size of each sub-region should be adjusted according to applications. Too large size may result in coarse version of blocks and cut down benefits of identifying important sub-regions, while too small size may result in identifying more important sub-regions. We should in general choose a block size that is in proportion to the volume size.

After partitioning, we quantify salient features of sub-regions. As previously introduced in Section 3, Shannon entropy can be used as an information content measure of a data set. Therefore, we can calculate the information entropy of every sub-region to estimate how many salient features it contains. Higher entropy value indicates more physical features. We define a normalized importance factor for each sub-region.

$$\omega_i^t = \frac{H(S_i^t)}{\max\limits_{i=1,\ldots,m} H(S_i^t)} \tag{5}$$

where S_j^t is the ith sub-region of the tth time step, m is total number of sub-regions. After calculating the importance factor, we sort them in descending order. A higher value of the importance factor indicates more information content, that is to say, scientists should pay more attention to the corresponding sub-region. Therefore, in the visualization process, we first sort importance factor of sub-regions, and then allocate time budget or rendering resources according to the distribution of importance factor. Sub-regions with higher importance factors are rendered in higher quality and priority while others are rendered in lower quality or even discarded.

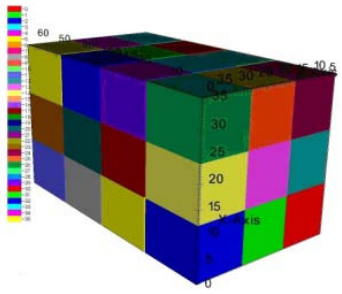


Fig. 2. Partitioning computational domain into sub-regions

V. EXPERIMENT RESULTS

In this section, we discuss the experimental results obtained with an implementation of our algorithm. All the following experiments were performed on a 2.66GHz Intel Core 2 Q9400 processor with 4 GB main memory, and an nVidia Geforce GTX260 graphics card with 896MB video memory.

5.1 Data Sets

The test data used in our experiments are listed in Table 1 and 2. The radiation diffusion dynamics and laser simulation data are from scientific simulations which were conducted by scientists at Institute of Applied Physics and Computational Mathematics (IAPCM). The radiation diffusion dynamics data (2D) simulates instability when temperature transmits to left boundary interface. The laser-plasma interaction data simulates electric field when laser incidents into plasma. The laser filament data simulates instability of high intensity laser propagation in plasma. The fuel and engine data are from volvis.org.

5.2 Time Step Selection

Firstly, we experimented with radiation diffusion dynamics data for time step selection. With the Algorithm 1, we selected 50 out of 1200 time steps with time cost of 0.51 second. Fig. 3(a) shows the KL-distances of original 1200 time steps. We can see that abnormal change was detected between time steps of about 500 and 900. Actually, temperature reached left boundary just at this time and some interesting physical phenomena took place. Fig.3(b) reveals the off-line marginal utility of each time step. It quantifies the utility of the current visualized time step and hence can be determined whether additional time steps are needed. Off-line marginal utility monotonically decreases as time passes because surprising information brought by new time steps decreases. Since the time step of about 900, there was nearly no new information from the followed time steps is helpful to understand the simulated process because temperature field tended to a stable state. Fig. 4(a) displays the distribution of the 50 time steps selected by the Algorithm 1. The selected time steps are mark with tag=1. For drawing a comparison, we also display the distribution of 50 uniformly selected time steps in Fig. 4(b). Our method yields a sequence of time steps which distribute densely when interesting physical phenomena take place and distribute sparsely when phenomena evolve stationarily and evenly. The results accord with our common sense. For a more intuitive view, we animated un-uniformly and uniformly selected 50 time steps, respectively. In Fig. 5, the upper lists a few of frames of un-uniformly selected time steps and the lower lists the same frames of uniformly selected time steps. From the animations, we can conclude that with the same number of time steps, un-uniformly selected time steps convey more information than uniformly selected ones.

TABLE 1. THE TIME-VARYING DATA SETS AND TIMINGS FOR TIME STEP SELECTION TEST.

data	volume dimension	time (s)
radiation diffusion dynamics	350×40×1200	0.51
laser-plasma interaction	80×80×201×2000	4.1

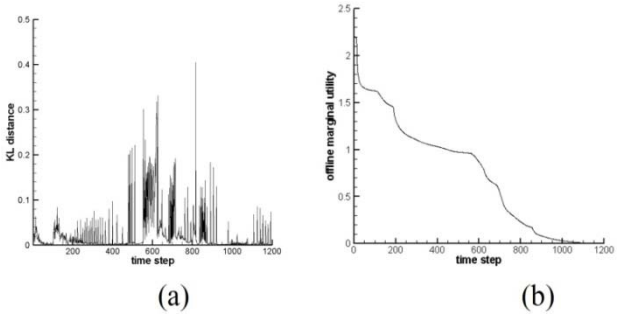


Fig. 3. Radiation diffusion dynamics data:(a) KL-distance between neighbor time steps. (b) Off-line marginal utility of each time steps.

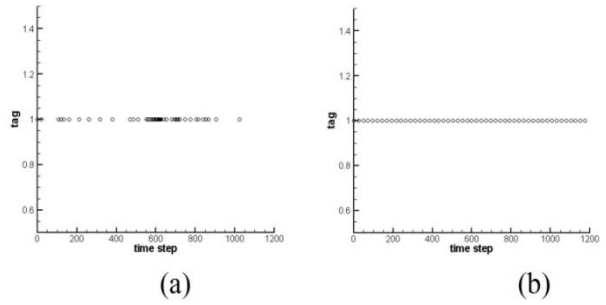


Fig. 4. 50 out of 1200 time steps are selected with radiation diffusion dynamics data (the selected time steps are marked with tag=1). (a) distribution of un-uniformly selected time steps. (b) distribution of uniformly selected time steps.

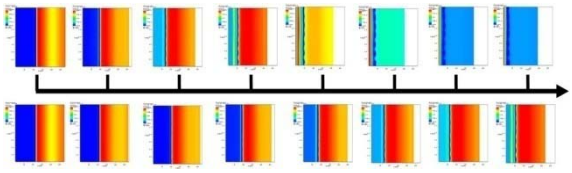


Fig. 5. Animation of selected time steps with radiation diffusion dynamics data. The upper is un-uniformly selected time steps associated with Fig.4 (a) and the lower is uniformly selected time steps associated with Fig. 4(b) The upper and the lower are the same frames. Left to right: 1th, 3th, 6th, 9th, 12th, 15th, 18th, 21th frame.

Similarly, we used the laser-plasma physics simulation data to illustrate the application of time step selection. As physical scientists didn't ascertain when to stop simulation process, they computed blindly 2000 time steps which were far more than needed. There was much redundant information in data and no need to visualize the all data sets. Using the Algorithm 1, we selected 10 out of 2000 time steps within 4.1 seconds. Fig. 6(a) shows the KL-distances of neighbor time steps. We can see that abnormal changes were detected at the time steps between about 0 and 900. Fig. 6(b) shows the off-line marginal utility of each time step. Similar with Fig. 3(b), the off-line marginal utility of laser-plasma physics simulation data also monotonically decreases as time passes because new time steps provide less and less new information. From Fig.6 (b), we can judge that since the time step of 1000, there was nearly no abnormal change took place. We guess that laser penetrated plasma and reach a stable state at about the time step of 1000. Fig. 7(b) shows volume visualization of the time step of 1000. It indicates laser has just reached the boundary. As physical scientists are interested in the process of penetration, time steps after 1000 are information redundancy and useless for scientists. Therefore, it is reasonable to truncate the time steps. Fig. 7(a) displays the distribution of the 10 time steps selected by Algorithm 1. The selected time steps are also mark with tag=1. The selected time steps distribute un-uniformly before laser reached boundary. These time steps are also more valuable than uniformly selected ones for scientists to understand physical discipline.

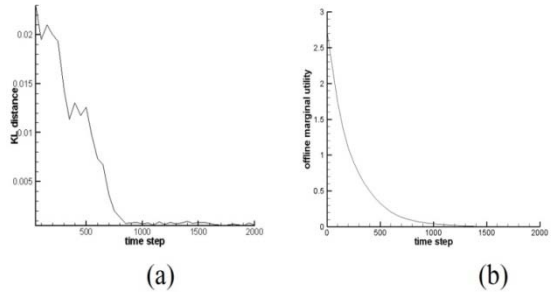


Fig. 6. Plasma physics simulation data:(a) KL-distance between neighbor time steps. (b) Off-line marginal utility of each time step.

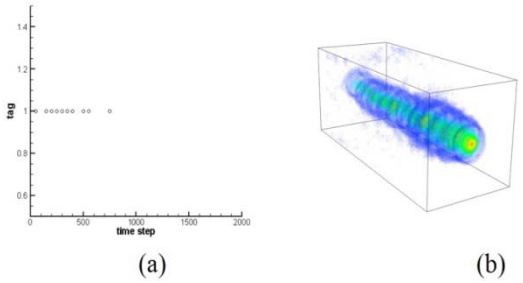


Fig. 7. (a) Distribution of un-uniformly selected time steps with plasma physics simulation data. 10 out of the 2000 time steps are selected. (b) Volume visualization of the time step 1000.

5.3 Sub-region Location

We tested important sub-region location with four data sets which listed in Table 2. Table 2 also lists the volume dimension, block dimension and the time cost.

After calculating importance factor of each block, we determine statistically significance from the distribution of block importance factor and set a valve value to locate important sub-regions. Fig. 8 shows distributions of block importance factor of four data sets. Obviously, most of blocks are unimportant and can be deemphasized. Red denotes importance factor value of selected sub-regions whose aggregate ratio is about 10%. All filtered blocks of each data set are shown in Fig. 9 with bounding box. Visualization results indicate that the proposed method can effectively capture sub-regions of interest.

TABLE 2. THE DATA SETS WITH THEIR PARAMETER SETTINGS AND TIMINGS FOR IMPORTANT SUB-REGIONS LOCATION TEST.

data	volume dimension	block dimension	time (s)
laser filament	512×512×1024	32×32×64	9.8
fuel	64×64×64	4×4×4	0.016
Laser-plasma	400×600×400	40×60×40	3.01
engine	256×256×256	16×16×16	0.25

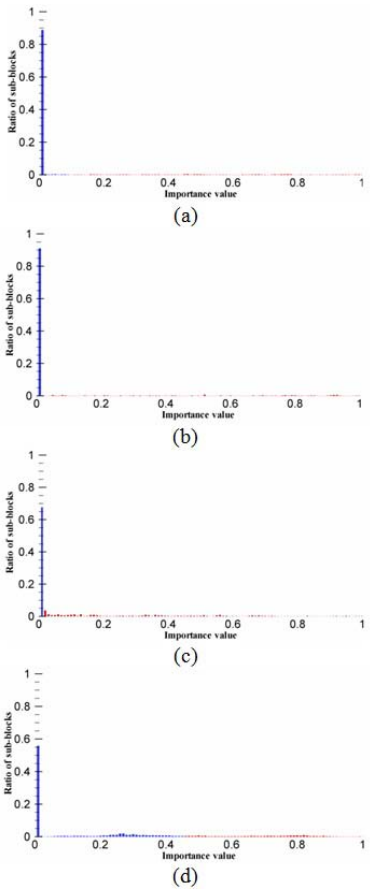


Fig. 8. Distribution of block importance factor. Red is importanc e factor value of selected sub-regions. (a) laser filament. (b) fuel. (c) laser-plasma interaction. (d) engine.

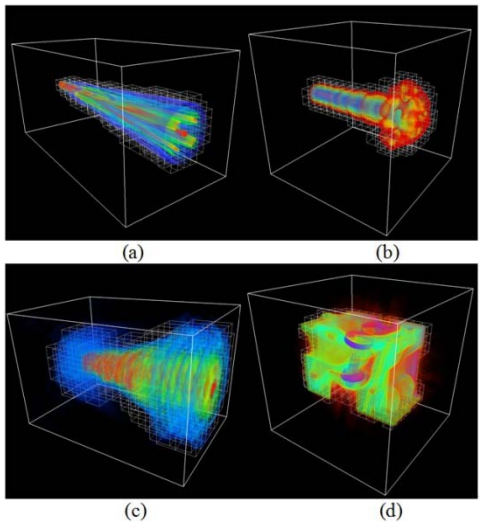


Fig. 9. Bounding box of selected sub-regions. (a) laser filament. (b) fuel. (c) laser-plasma interaction. (d) engine.

VI. DISCUSSION

As previous described, the proposed un-uniform algorithm is more reasonable than uniform method for selecting the same number of time steps. Consider an extreme situation when there is a stable feature that only moves in space, the un-uniform selection algorithm degenerates to uniform selection method. The important sub-region location method utilizes intrinsic information content rather than explicit feature description. Thus, the proposed method is general and useful for extensive applications.

The proposed method is low time cost. The main time-consuming computation of the method is to construct probability distribution which is O (n) complexity (e.g. histogram). Further more, our work targets time-varying scientific simulation data where distributed data are usual. During distributed processing, it is not necessary to communicate massive original data. We only need to compute local probability distributions respectively and communicate them. Therefore, our method is suitable to parallel computing environment.

With information assisted visualization techniques, we can accelerate visualization process and enable scientists to examine large scale time-varying data within limited time cost or computing resources. Our work also provides a new potential direction to reduce data storage space, to optimize computing resources, and to maximize the scientist's ability to understand time-varying data.

VII. CONCLUSION

In this paper, we propose an information assisted visualization technique to extract and visualize important physical features in time-varying scientific data. We have utilized KL-distance and off-line marginal utility to select most representative time steps and defined importance factor to locate sub-regions which are most unusual in information content. With the extracted time steps or sub-regions from the overall data, we are able to present most informative aspect of the data. Favorable experiment results obtained with scientific data validate the effectiveness of our method.

ACKNOWLEDGEMENT

This research was supported by National Natural Science Foundation of China (61033009), National Basic Key Research Special Fund (2011CB309702), the Science and Technology Funds CAEP under grant 2010B0403057, Chinese National Science Foundation under grant 61003083

REFERENCES

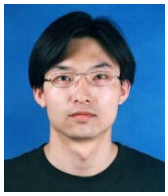
[1] N. D. B. Bruce: Features that draw visual attention: an information theoretic perspective. *Neurocomputing*, volume 65-66, pages 125-133, 2005.

[2] U. D. Bordoloi, H.-W. Shen: View Selection for volume rendering, In Proc. of *IEEE Visualization'05*, pages 487-494, 2005.

[3] P. Barford, A. Bestavros, J. Byers, M. Crovella: On the marginal utility of network topology measurements. *In Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, pages 5-17, San Francisco, California, 2001.

[4] S. Boltz, E. Debreuve, M. Barlaud: High-dimensional statistical distance for region-of-interest tracking. *IEEE Trans. On Image Processing*, 18(6): 1266-1283, 2009.
[5] J. Caban, A. Joshi, P. Rheingans: Texture-based feature tracking for effective time-varying data visualization. *IEEE Trans. on Visualization and Computer Graphics*, 13(6):1472-1479, 2007.
[6] Chen M., Jaenicke H. An information-theoretic framework for visualization. *IEEE Trans. on Visualization and Computer Graphics*, 16(6):1206-1215, 2010.
[7] T. Dasu, S. Krishnan, S. Venkatasubramanian, K. Yi: An information-theoretic approach to detecting changes in multi-dimensional data streams. *In Proceedings of the 38th Symposium on the Interface of Statistics, Computing Science, and Applications (Interface'06*).
[8] N. Fout, K.-L. Ma, J. Ahrens: Time-varying, multivariate volume data reduction. *In Proc. of ACM Symposium on Applied Computing*, pages 1224-1230, New York, 2005.
[9] M. Feixas, M. Sbert, F. Gonzalez: A unified information-theoretic framework for viewpoint selection and mesh saliency. A*CM Trans. On Applied Perception*, 6(1), 2009.
[10] G. Ji, H.-W. Shen: Dynamic view selection for time-varying volumes. *IEEE Trans. on Visualization and Computer Graphics*, 12(5):1109-1116, 2006.
[11] D. Jing, W. Wang, Y. Zhao: General correlation coefficient between variables based on mutual information. Journal of Sichuan University, Vol.34, No.3,2002.
[12] C. Kamath: Scientific data mining: a practical perspective. *SIAM Press, Philadelphia*, 2009.
[13] J. Lin: Divergence measures based on the Shannon entropy. *IEEE Trans. on Information Theory*. 37(1): 145-151. 1991.
[14] K.-L. Ma: Large scale data visualization. *IEEE Computer Graphics and Applications*, 21(4):22-23, 2001.
[15] A. Pescape: Entropy-based reduction of traffic data. *IEEE Communications Letter*, 11(2): 191-193, 2007.
[16] C. Shannon: A mathematical theory of communication, Bell Systems Technical J., volume 47, pages 143-157, 1948.
[17] H.-W. Shen: Visualization of large scale time-varying scientific data. In Proceedings of *SciDAC 2006*, Journal of Physics: Conference Series, 46(1):535-544, 2006.
[18] H.-W. Shen, L.-J. Chang, K.-L. Ma: Time-varying volume rendering using a time-space partitioning tree. *IEEE Visualization '99*, pages 371-377, San Francisco, California, October, 1999.
[19] X. Song, M. Wu, C. Jermaine, S. Ranka: Statistical change detection for multi-dimensional data, In Proc. Of *International Conference on Knowledge Discovery and Data Minin*g, pages 667-676, San Jose, California, 2007.
[20] I. Viola, M. Feixas, M. Sbert, M. E. Groller: Importance-driven focus of attention. *IEEE Trans. on Visualization and Computer Graphics*, 12(5):933-940, 2006.
[21] C. Wang, H. Yu, K.-L. Ma: Importance-driven time-varying data visualization. *IEEE Trans. on Visualization and Computer Graphics*, 14(10):1547-1554, 2008.
[22] C. Wang, H. Yu, K.-L. Ma: Application-driven compression for visualizing large-scale time-varying volume data. *IEEE Computer Graphics and Applications*, 30(1):59-69, 2010.
[23] C. Wang K.-L., Ma: Information and knowledge assisted analysis and visualization of large-scale data, *Ultrascale Visualization'08*, pages 1-8, 2008.
[24] Y. G. Wu: Region of interest image indexing system by DCT and entropy. *International Journal on Graphics, Vision and Image Processing*, 6(4): 235-244, 2006.
[25] L. Xu, T. Lee, H.-W. Shen: An information-theoretic framework for flow visualization. *IEEE Trans. on Visualization and Computer Graphics*, 16(6):1216-1224, 2010.

Guoqing Wu received the B.Sc. degree in information and computational science from Xi'an Jiaotong University, China in 2003 and the M.S. and Ph.D. degrees in computer science from China Academy of Engineering Physics in 2006 and 2009 respectively. He is currently an assistant researcher at High Performance Computing Center, Institute of Applied Physics and Computational Mathematics, Bei Jing. His research interests include information visualization and scientific data mining.



Yi Cao received the B.Sc. degree in computer science from Lanzhou University, China in 1998 and the M.S. degree in computer science from China Academy of Engineering Physics in 2007. He is currently an associate researcher at High Performance Computing Center, Institute of Applied Physics and Computational Mathematics, Bei Jing. His research interest is scientific visualization.

Junping Yin received the B.Sc. degree in applied mathematics from Northeast Normal University, China in 2002 and the M.S. and Ph.D. degrees in applied mathematics from Ximen University in 2008. He is currently an assistant researcher at High Performance Computing Center, Institute of Applied Physics and Computational Mathematics, Bei Jing. His research interest is statistics.