Multiple-Scale Visualization of Large Data Based on Hierarchical Clustering

Bao T. Nguyen, Emanuele Olivetti, and Paolo Avesani

***Abstract*—Nowadays, the large datasets become more and more common. However, traditional visualization techniques, which although allow to visually analyze and explore data, can not scale well with the large one. This restrains the ability of detecting, recognizing and classifying phenomena of interest, such as patterns, clusters, trends, etc. This Paper proposes a method for interactive multi-resolution visualization to overcome the problem of traditional visualization techniques when working with a large dataset via hierarchical clustering. Based on hierarchical clustering, users can not only examine the dataset at different levels of detail, but also can explore many regions of interest. The basic idea underlying this method is to choose multiple scales from the hierarchical tree for representing the data at different levels of abstraction, which creates an easy environment for interactive exploration without re-run the clustering algorithm. Moreover, we also define a criterion for evaluating the multiple scale representation based on the concept of split factor. An experiment of applying the proposed method into the task of interactive visualization in a clinical case study of the large dMRI (diffusion magnetic resonance imaging) data is also carried out. The results show that our proposed method efficiently provides a friendly tool for visualization the large data.**

***Index Terms*—Visualization, hierarchical clustering, machine learning, tract segmentation, diffusion MRI.**

I. Introduction

To support human in analyzing and exploring large data, it is an important task to graphically present the data [1]. Users, in one side, have a requirement of looking at complex and intricate data to find out some facts or trends that are not easy to find. On the other side, they want to explore data in details to examine each data points. In fact, the overall premise is that users have a deeper understanding about their data when they interact with the presented information and view it at different levels of abstraction [2]. During the last two decades, many interactive visualization techniques and system have been emerged [1], [3], [4]. As large data sets become more and more common, with the size over 1K, it has been clear that most of the current visualization approaches lose their effectiveness due to they have no ability to visualize and manage the large number of data points simultaneously. In such scenario, clustering is considered a suitable method for understanding and exploring large data [5], [6].

However, clustering usually results in one partition of the data, and this leads to a dramatic drawback that the validation process is not straightforward due to the lack of ground truth data [7]. One solution for this is hierarchical clustering [8], which organizes data in an intuitive and interpretable structure, namely *dendrogram*, not only one partition as the traditional clustering methods. Such structure allows users to explore in a simple way the clusters and the relationships between instances, and leads to many applications for visualization [9]-[11]. Nevertheless, when dealing practically with a large size dendrogram, it becomes difficult since the number of nodes grows exponentially with the depth of the tree and makes users lose the overview of the whole dataset. To deal with a large size dendrogram, many approaches have been suggested [6], [10], [12]. However, these methods are either display at one time only a sub-part of the structure [10], [12], or display whole dendrogram but rely on other clustering technique [6].

In this work, we propose a method of offering a complete and interactive visualization of the large data based on hierarchical clustering. Our method allows users to apply their perceptual abilities to make sense of data. The core of the problem is to obtain the multiple scales representation large data, in order to comply with the requirements of human interactive visualization. The proposed solution combines three steps. First, the dendrogram would be created by running the hierarchical clustering. Second, the goodness function is used as a measurement to select the most relevant scales for representing the dendrogram (it is an extension of the "relevant function", proposed in [13]). Lastly, we evaluate the multiple scales based on a statistical criteria, called *split factor*.

Moreover, we conceive an experiment of applying our method in a clinical case study of dMRI data. Recently, from dMRI data, tracking algorithms [14], [15] allow to reconstruct the 3D pathways of axons within the white matter as a set of streamlines, called *tractography*. A *streamline* is a vectorial representation of thousands of neuronal axons expressing structural connectivity, and tractography is a set of N streamlines ($N \sim 3 \times 10^5$ usually). It is an important task of segmentation the tractography into some real anatomical structures of interest, such as cortinal spinal tract [16], [17] involving to the amiotrophic (ALS) disease. In this experiment, we conceive a novel computer-assisted interactive segmentation process based on our method of multiple scale representation of the large tractography.

The paper is organized as follows. Section II formally introduces the problem of multiple scales for representation large data. After that, Section III describes the detail of the method for selecting multiple scales. The evaluating the goodness of the representation is presented in the next Section IV. In Section V, we describe an experiment of applying the proposed solution in the context of the tractgoraphy segmentation, provide figures to evaluate the viability of the proposed solution. We conclude with a

Manuscript received August 20, 2013; revised November 13, 2013.

The authors are with the Bruno Kessler Foundation (FBK) and Trento University, Trento, 38122, Italy (e-mail: tbnguyen@fbk.eu, olivetti@fbk.eu, avesani@fbk.eu).

summary of our contribution and open areas for future work in the last Section VI.

II. Problem Statement

In this part, after introducing the hierarchical clustering we formally describe the problem of multiple scale representation.

A. Hierarchical Clustering

Given a set of input patterns denoted as $X = \{x_1, ..., x_j, ..., x_N\}$ where each data point $x_j = (x_{j1}, x_{j2}, ..., x_{jd})^T \in R^d$ and each measure x_{ji} is said to be a feature (attribute, dimension, or variable). A hierarchical tree (or dendrogram) of X is defined as following:

Definition 1: A *hierarchical tree H* of an N-object set $X=\{x_1, ..., x_N\}$ is a collection of Q partitions on X: $H = \{P_0, ..., P_Q\}$, with $Q \le N$, such that $P_0 = X$ and $C_i \in P_m$, $C_j \in P_l$, $m > l$ imply $C_i \subseteq C_j$ or $C_i \cap C_j = \varnothing$, for all $i, j \ne i$, and $m, l \in [1, Q]$.

The hierarchical clustering algorithm [8] builds nested clusters by merging them successively, and this hierarchy of clusters represented as a tree/dendrogram. The root of the tree is the unique cluster that gathers all the samples, the leaves being the clusters with only one sample. It produces a structure of clusters of X that is more informative than the unstructured set of clusters returned by flat clustering. This characteristic meets the requirement of creating multiple scales of one original dataset $X = \{x_1, ..., x_N\}$ without re-running the clustering algorithm again. Moreover, it leads to the capability of visualizing X in many levels of abstraction, and the users can browse the value of level from 1 to N, to see the clusters immediately.

Hierarchical clustering algorithms are either top down or bottom up. Bottom-up algorithms treat each streamline as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all tracts. Bottom-up hierarchical clustering is therefore called Hierarchical Agglomerative Clustering (HAC). Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual streamlines are reached [8].

B. Multiple scales for Visualization

The hierarchical tree H structures and presents dataset X at different levels of abstraction. A non-leaf cluster is composed of all its child clusters, while a leaf cluster contains only a single data item. The collection of all leaf-clusters presents exactly every data items x_i of X, while the root is a cluster containing whole dataset X as one single node of the tree.

Definition 2: Each cluster C_i (node) of the tree H, let $s(C_i)$ be *the level of detail* of that cluster. This measurement $s(C_i)$ satisfies the following criteria: if C_i is an ancestor of C_j, then $s(C_i) \ge s(C)$.

There are many properties of a cluster which could be used to measure $s(C_i)$. Among these, two common uses are the radius of a cluster (maximum distance between all pair samples of cluster C_i: $s(C_i) = \max\{d(x_a, x_b)\}$, with $\forall x_a, x_b \in C_i$ and $x_a \ne x_b$; and the and the hierarchical level of C_i in the tree H: $s(C_i) = \frac{height(C_i)}{h}$, where height($C_i$) is the height of the cluster C_i, and h is the height of the tree H [1].

Definition 3*:* The range of scale of a hierarchical tree H is $[s_{min}, s_{max}]$, where $s_{min} = \min\{s(C_i)\}$, with $\forall C_i \in H$; and $s_{max} = \max\{s(C_i)\}$, with $\forall C_i \in H$.

Depending on which property is used to measure the level of detail s_i, the value of s_{max} and s_{min} would be different. In the case of using the hierarchical level, the scale range is from [0, 1], where $s_{min} = 0$ corresponds to the leaf with zero height, and $s_{max} = 1$ is at the root of the tree H. However, in the case of using cluster radius, there is no guarantee that $s_{min} = 0$ and $s_{max} = 1$.

Definition 4: A cut L of a hierarchical tree H at a given scale $w \in [s_{min}, s_{max}]$ is $L(w)$:

$$L(w) = \left\{ C_i \,\middle|\, s(C_i) \le w \wedge s(parent(C_i)) > w \right\} \quad (1)$$

where parent(C_i) is the direct parent node of the cluster C_i. In general, $L(w)$ is a partition of X , denoting a subset of the tree H. The cut at s_{min}, $L(s_{min})$ is a set of all leaf clusters, while the $L(s_{max})$ is a single cluster representing the whole dataset X. Intuitively, $L(w)$ changes smoothly with the variance of the scale parameter w, which serves as the abstraction level of the dataset X. It could be imagined that $L(w)$ is a cut across a vertically oriented hierarchical tree H that satisfies criteria: $L(w)$ intersects each path of the tree H, from the root to the leaf, only exactly at one point. The cutting point would depend on the value of parameter w. It should close to the root of the tree H when w is high, and reversely. Moreover, the cut can be horizontal or un-horizontal (like zigzag) as long as for each path from the root to the leaf of the tree H, there is only one crossing with $L(w)$. It is an open approach for cutting the tree comparing with the traditional one which only accepts the horizontal cut.

Definition 5: Let P and Q be two partitions of dataset X, $P = \{C_1^P, \ldots, C_l^P\}$ and $Q = \{C_1^Q, \ldots, C_m^Q\}$. Partition P is *nested* in partition Q, denoted as $P \prec Q$, if and only if:

$$P \prec Q \leftrightarrow \forall C_i^Q \in Q, \exists C_{i_1}^P, \ldots, C_{i_k}^P \in P : C_i^Q = \bigcup_{t=1}^{k} C_{i_t}^P \quad (2)$$

Definition 6*:* Given the scale range $[s_{min}, s_{max}]$ of a tree H, the ***multiple scales representation*** for the tree H is an ordered set of k scale values from $[s_{min}, s_{max}]$: $B = \{b_1, b_2, \ldots, b_k\}$, with $b_i \in [s_{min}, s_{max}]$, $\forall i \in [1, k]$, where k is the order of the set B, which satisfies the following condition:

$$\forall i \in [1, k-1]: L(b_i) \mathbf{p} L(b_{i+1}) \quad (3)$$

Multiple scale representation problem: Given a hierarchical tree H on a dataset X , with the scale range $[s_{min}, s_{max}]$. How to choose the multiple scales representing for the tree H: $B = \{b_1, b_2, \ldots, b_k\}$, with $b_i \in [s_{min}, s_{max}]$, $\forall i \in [1, k]$? It is an NP-problem, and there is no general solution for it. Usually, it is chosen that $b_1 = s_{min}$, where the whole elements of X are presented, and $b_k = s_{max}$, which corresponds to only one virtual representation of X. However, the value of k is an open question and totally depends on the application. In the next section, we will discuss about how to define the k value and also how to select each b_i from $[s_{min}, s_{max}]$.

III. Methods

In this part, we present a simple and efficient method to determine the multiple scales $B = \{b_1, b_2, \ldots, b_k\}$, with $b_i \in [s_{\min}, s_{\max}]$, $\forall i \in [1, k]$, where $[s_{\min}, s_{\max}]$ is the range scale of the hierarchical tree H, constructed from dataset X. Moreover, the multiple scales B have to satisfy the condition in Definition 6.

Definition 7: Given a cluster C_i in a hierarchical tree H, with range scale $[s_{\min}, s_{\max}]$, the pairwise $\left({}^{C_i}_{min}, {}^{C_i}_{max} \right)$ is defined as:

$$\begin{aligned} {}^{C_i}_{min} &= min \left\{ w_j \mid w_j \in [s_{min}, s_{max}] \ \wedge \ C_i \in L(w_j) \right\} \\ {}^{C_i}_{max} &= max \left\{ w_j \mid w_j \in [s_{min}, s_{max}] \wedge C_i \in L(w_j) \right\} \end{aligned} \quad (4)$$

Intuitively, $\left({}^{C_i}_{max} - {}^{C_i}_{min} \right)$ are two scale factors at which the cluster C_i appears and disappears from the tree H. It is considered that the good clusters would be presented for a wide range of scale factors. Thus, the goodness of a cluster could be measured as $\left({}^{C_i}_{max} - {}^{C_i}_{min} \right)$, and the best scale representing C_i be as $= \frac{{}^{C_i}_{max} - {}^{C_i}_{min}}{2}$. These two characters can be included in the following function:

Definition 8: The goodness function $R(C_i)$ of a cluster C_i at the scale w is:

$$R_w(C_i) = \frac{a^{C_i}_{\max} - a^{C_i}_{\min}}{2} + \frac{2(a^{C_i}_{\max} - w)(w - a^{C_i}_{\min})}{a^{C_i}_{\max} - a^{C_i}_{\min}} \quad (5)$$

Definition 9: Given a scale $w \in [s_{\min}, s_{\max}]$, the goodness function $R(w)$ of a scale w is:

$$R(w) = \frac{1}{N} \sum_{C_i \in L(w)} |C_i| R_w(C_i) \quad (6)$$

A plot line of the $R(w)$ function can be found in the Fig. 1. Obviously, $R(w)$ is a quadratic function of w, and can be used for determining the scale factors corresponding to the good clusters. By focusing on the local maxima of *R(w)*, we can estimate good scales for representing the tree H, and thus getting the $B = \{b_1, b_2, \ldots, b_k\}$, with $b_i \in [s_{\min}, s_{\max}]$, $\forall i \in [1, k]$. In another way, by setting the first derivation of *R(w)* from w to zero, and we could arrive a set of multiple scales B.

$$B = \left\{ b_i \mid b_i \in [s_{\min}, s_{\max}] \wedge \frac{\partial(R(b_i))}{\partial(b_i)} = 0 \right\} \quad (7)$$

The most difficult task is to compute the pairwise $\left({}^{C_i}_{min}, {}^{C_i}_{max} \right)$ for each cluster C_i H. Pascal *et al.* in [13] proposed a method to calculate $\left({}^{C_i}_{min}, {}^{C_i}_{max} \right)$ based on the concept of relevant community. However, the proposed procedure is computational cost, and the complexity is between $O(n \log n)$ and $O(n^2)$ with an average value in $O(n*\sqrt{n})$. As the meanwhile, the hierarchical order of the tree H provides a good hint about the scales where each cluster appears or disappears. By exploring this information, we suggest a more easy and efficient way with the complexity O(1): ${}^{C_i}_{min} = s(C_i)$ and ${}^{C_i}_{max} = s(parent(C_i))$, where $s(C_k)$ is the level of detail of the cluster C_k (as the Definition 2). Obviously, the suggested way to compute $\left({}^{C_i}_{min}, {}^{C_i}_{max} \right)$ intuitively satisfies the Definition 7.

IV. Evaluation

In this part we describe criteria for evaluating the multiple scale representation $B = \{b_1, b_2, \ldots, b_k\}$, of the dataset X. Due to the limitation of the screen size, it is the real fact that, at a certain time, users can only exam about the total of 50 (l_1) clusters which are currently displaying on the screen. Among of the visible clusters, the users usually select around 15~25 (l_2) clusters to explore or exam the detail [18]. Driven from that, we propose a method to evaluate the represented multi-scale set B based on a quantitative measure, called *split factor*, as following.

Definition 10: *Split factor* ξ of a cluster $C_i \in H$ to a scale $s \in [s_{\min}, s_{\max}]$ is $\xi(C_i, s)$

$$x(C_i, s) = card\left(P(C_i, s)\right) \quad (8)$$

where $P(C_i, s) = \{C_j \mid (C_j \in H) \wedge (s(C_j) = w) \wedge (C_j \subseteq C_i)\}$ Based on the split factor of a cluster, we can define the split factor for a set of clusters or a partition as:

Definition 11: Split factor ξ of a set of clusters $P = \{C_1, C_2, ..., C_m\} \subseteq H$ to a scale $s \in [s_{\min}, s_{\max}]$ is $\xi(P, s)$

$$\xi(P, s) = \sum_{C_i \in P} \xi(C_i, s) \quad (9)$$

The quantitative measure of split factor can be used to evaluate the multiple scales representation $B = \{b_1, \ldots, b_k\}$ by computing the split factor for all of the cut scale $L(b_i)$, $\forall i \in [1, ..., k]$. However, as already stated before, at a certain time, the users usually only explore or exam about (λ_2) currently visualized clusters. Due to this, we should not calculate the split factor for the whole partition $L(b_i)$, instead of that, only a subset of $L(b_i)$ with the order of λ_2 should be used. At each specific scale $b_i \in B$, called $S_{(b_i, \lambda_2)}$ is a Gaussian distribution subset of the cut H at scale b_i, $L(b_i)$, with the order of λ_2:

$$S_{(b_i, \lambda_2)} = \left\{ C_1, \ldots, C_{\lambda_2} \right\} with \quad C_j \in L(b_i), \forall j \in [1, \ldots, \lambda_2] \quad (10)$$

The evaluation procedure can be done as the following definition:

Definition 12: The set of scales $B = \{b_1, b_2, \ldots, b_k\}$ is called *the best scales* for representation of the tree H, given λ_1 and λ_2, if the following condition satisfies

$$\forall b_i \in B: \quad \lambda_1 - \Delta \le \xi\left(S_{(b_i, \lambda_2)}, b_{i-1} \right) \le \lambda_1 + \Delta \quad (11)$$

In the case of b_1 the split factor is computed to the leaf, $\xi(S_{(b_1, l_2)}, 0)$, and Δ is a non-negative scalar. The smaller the value of Δ is, the more strict the degree of the best scales is.

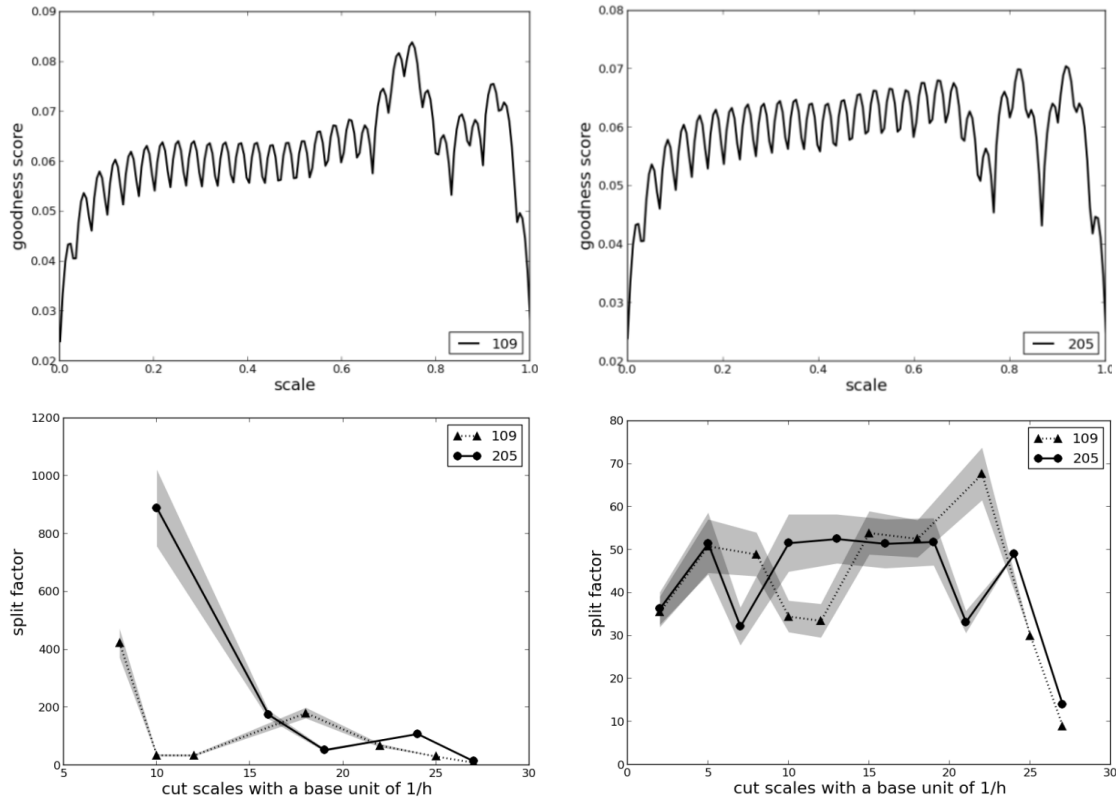

Fig. 1. Top line: goodness score of subject 109 and 205 from ALS dataset at different scales. The scales having local maximum value of goodness score are chosen as the representation of dMRI data. Bottom line: split factor before (left) and after (right) adding heuristic constrain. After applying heuristic constrain, the mean split factors are around the expected value $\lambda 1 = 50$ (with $\Delta = 15$).

V. Experiments

In the following we briefly describe our experiment for visualizing the real large tractography to validate the proposed approach. The evaluation based on split factor and one heuristic trick to improve the visualization result are also presented.

A. DMRI and Tractography Segmentation

Let the polyline $s = \{x_1, ..., x_N\}$, where $x_i \in R^3$, be a streamline reconstructed from dMRI data by deterministic tractography algorithms [14]. Let the tractography $T = \{s_1, ..., s_N\}$ be defined as a set of N streamlines. Our experiment is motivated by a clinical research hypothesis about the characterization of the amiotrophic (ALS) disease, which is known to be affected by the corticospinal tract (CST) [16], [17]. The first task is to segment the CTS from the full brain tractography *T*.

In spite that recently there is an increasing literature in automatic tractography segmentation using machine learning techniques [19], [20], applications in the clinical domain rely on manual segmentation. The manual segmentation process consumes a lot of time and effort due to the large number of streamlines, in the order of 3×10^5, which make it intrinsically difficult both to inspect and to unfold the anatomical structures. Moreover, it is claimed that there is a lack of software tools to support and to simply this segmentation process [18]. In this experiment, we conceive a novel computer-assisted interactive process based on the method of multiple scales for representation the large tractography described in Section III. After computing the set of multiple scales $B = \{b_1, ..., b_k\}$, our tool first displays *T* as the cut at b_1, $L(b_1)$; and let user select some of clusters to identify a superset of the streamlines of interest. This superset is then to be displayed at the next scale and again the user is requested to select the relevant clusters. The process of re-display and manual selection is iterated until the remaining streamlines faithfully represent the desired anatomical structure of interest.

ALS dataset: the data we used in this experiment is recorded with a 3T scanner at Utah Brain Institute. It consisted the recordings of 12 ALS patients and 12 healthy controls; 64 (+1, i.e. $b = 0$) gradients; *b*-value= 1000; anatomical scan ($2 \times 2 \times 2 mm^3$). We reconstruct the streamlines using EuDX, a deterministic tracking algorithm [21] from the DiPy library.[1]

Dissimilarity representation: due to the fact that each streamline has different length and different number of points we need to find a representation $\varnothing$ of streamline in a vectorial space, by mapping a streamline s from its original space T to a vector of R^p - $\Theta : T \to R^p$, where p is the dimension of the new space. One suggestion for this is the dissimilarity representation [22]. It is a lossy Euclidean embedding algorithm was previously proposed in [23] for streamlines. The dissimilarity representation is defined as $\Theta_{\Pi}^{d}(X) = \left[d(X, X_1), \mathbf{K}, d(X, X_p) \right]$, where d is a distance function between streamlines, and $\Pi = \{X_1, ..., X_p\} \subset X$ is a set of p streamlines, called *prototypes* (detail in [23], [24]).

By applying the hierarchical clustering algorithm (Section II-A) on the dissimilarity approximation, the hierarchical tree

[1] http://www.dipy.org

H of the tractography T could be created.

B. Multiple scales for Representation

As the measurement for computing the level of detail of a cluster $s(C_i)$, we use the height of the cluster C_i within the hierarchical tree H. The reason is that this measurement leads to continuous and thus provides smooth transitions on our hierarchical display. Let h be the height of hierarchical tree H: $h = height(H)$, at the leaf C_{leaf} of H, the height is in the order of zero, thus $s_{\min} = 0$. In the similar way, $s_{\max} = 1$ because at the root C_{root} of the tree H, $height(C_{root}) = h$. The range scale of H is $[s_{\min}, s_{\max}] = [0, 1]$, and $\forall\, C_i \in H$, $s(C_i) = height(C_i)$, where $height(C_i)$ is the height of the cluster C_i [1]. Intuitively, this measurement satisfies the condition of Definition 2 about the level of detail in, because if C_i is an ancestor of C_j, then $height(C_i) \geq height(C_j)$ and thus, $s(C_i) \geq s(C_j)$.

Looking at the local maxima of the goodness score as in the Definition 9, we can estimate the most relevant scale factors for representing the tree H, and thus get the multiple scale representation $B = \{b_1, ..., b_k\}$. The top line of the Fig. 1 shows the plot lines of two goodness scores of subject 109 (left) and control 205 (right) from ALS dataset. For example with subject 109 (left) the multiple scale representation B_1 could be concluded as $B_1 = \{8/h_1, 10/h_1, 12/h_1, 18/h_1, 22/h_1, 25/h_1, 27/h_1\}$, where h_1 is the height of the hierarchical tree H of subject 109: $h_1 = height(H_{109})$. Similarly, with control 205, $BB = \{ 10/h_2, 16/h_2, 19/h_2, 24/h_2, 27/h_2\}$, where $h_2 = height(H_{205})$. Taking into account that b_i should be chosen from the small scale factor to the large one in order to satisfy the condition of an ordered set in the Definition 6, of which the underlying idea is to make sure a continuous and smooth order of visualization when users switch among these levels. This experiment exams the ability of our proposed method to compute the multiple scales representation for a large data. Only the two samples of results are presented, we also run on other subjects and get the equivalent multiple scale representation for each of them.

We are now in the state of being ready to evaluate the multiple scale representation $B = \{b_1, ..., b_k\}$. Based on the Definition 12 about the goodness of a scale factor $b_i \in B$, we implemented a program for evaluating the best scales representation B with $\lambda_1 = 50$ and $\lambda_2 = 15$. At the bottom left of the Fig. 1, we plot the mean goodness score of each multiple scale representation for subject 109 and control 205 from ALS dataset, together the standard derivation of 20 iterations. Note that on the horizontal axis, the cut scales l represented on the figure is the index of the corresponding real scale l/h.

Except for the first chosen scale b_1, almost other scale $b_i \in B$, $i \neq 1$, the split factor $\xi(S_{(b_i,\lambda_2)}, b_{i+1})$ satisfies the condition in Equation 11. Note that from the leaf (scale 0) the goodness score increases linearly, reaches the peak at scale b_1, and then gets a fluctuating variety. It shows that the split factor of the cut at b_1 to the leaf (scale 0), $\xi(S_{(b_1,\lambda_2)}, 0)$, is usually very large comparing with other split factors. However, in the point view of visualization, all the split factor $\xi(S_{(b_i,\lambda_2)}, b_{i+1})$ should be around λ_1. Due to this, for the sake of better representation, we add an heuristic constrain to the chosen scale set: the distance between two closest scales b_i and b_{i-1} should not exceed a threshold δ: $d(b_i, b_{i-1}) \leq \delta$ (in the case of b_1, the distance with the leaf, $d(b_1, 0)$, is used). The results after adding constrain are showed in the Fig. 1 -bottom right (with $\delta = 4/h$), and the mean split factor is closer to $\lambda_1 = 50$ (with $\Delta = 15$) than before adding heuristic constraint. Moreover, we also run on many subjects from ALS dataset, the sizes of which vary from 200K to 300K. These results demonstrate that our proposed method of choosing multiple scales for visualization large data is efficient and robust to the size of the large data.

VI. Conclusion

In this paper, we presented a method for addressing the problem faced when attempting to interactively visualize a large dataset. The core principle behinds the framework was to choose *multiple scales for representing* the data from the hierarchical clustering. Moreover, we also proposed a function to evaluate the goodness of each chosen scale based on the concept of *split factor*. We instantiate this framework with an application of building the interactive visualization large dMRI data in the procedure of tractogaphy segmentation, and provide concrete result on it performance. Experiments have shown that our method provide a significant improvement for visualization the large data at different scales, which verifies the effectiveness of the interactive hierarchical visualization. Besides, we are convinced that this method can be easily integrated to any current display techniques without having to vary the data or the interactive exploration tool.

As mention is the Section III, the level of detail of each cluster, $s(C_i)$, can be computed based on radius or height [1]. In this paper we choose the multiple scales only based on the height of cluster. The same job but based on the radius needs to be investigated. Moreover, this work is a part of a going research project focusing on computer-aided tractography segmentation, where machine learning techniques are used to assist medical practitioners to do the segmentation task more easily, flexibly and effectively. In the future, we want to further improve the interactive segmentation tool by providing the function of adding or eliminating data points x into or from the current dataset X, and updating the visualization result without re-run the clustering algorithm.

References

[1] J. Yang, M. O. Ward, and E. A. Rundensteiner, "Interactive hierarchical displays: a general framework for visualization and exploration of large multivariate data sets," *Computers & Graphics*, vol. 27, no. 2, pp. 265–283, Apr. 2003.

[2] J. C. Roberts, "State of the art: coordinated & multiple views in exploratory visualization," in *Proc. 5th Intl Conference on Coordinated and Multiple Views in Exploratory Visualization*, Washington, DC, USA, July 2007, pp. 61–71. [Online]. Available: http://dx.doi.org/10.1109/cmv.2007.20

[3] I. Stroe, E. Rundensteiner, and M. Ward, "Scalable visual hierarchy exploration," in *Database and Expert Systems Applications, ser. Lecture Notes in Computer Science*, M. Ibrahim, J. Kung, and N. Revell, Eds. Springer Berlin Heidelberg, 2000, vol. 1873, pp. 784–793.

[4] Y. H. Fua, M. O. Ward, and E. A. Rundensteiner, "Structure-basedbrushes: a mechanism for navigating hierarchically organized data and information spaces," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 2, pp.150–159, Apr. 2000.

[5] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping Multidimensional Data*, J. Kogan, C. Nicholas, and Teboulle, Eds. Berlin/Heidelberg: Springer Berlin Heidelberg, 2006, ch. 2, pp. 25–71.

[6] G. Bisson and R. Blanch, "Improving Visualization of Large Hierarchical Clustering," in *Proc. 16th Intl Conference on Information Visualisation*, Jul. 2012, pp. 220–228.
[7] L. Candillier, I. Tellier, F. Torre, and O. Bousquet, "Cascade evaluation of clustering algorithms," in *Machine Learning: ECML 2006, ser. Lecture Notes in Computer Science*, J. Furnkranz, T. Scheffer, and M.Spiliopoulou, Eds. Springer Berlin Heidelberg, 2006, vol. 4212, pp.574–581
[8] S. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, Sep. 1967.
[9] J. Heard, W. Kaufmann, and X. Guan, "A novel method for large tree visualization." *Bioinformatics*, vol. 25, no. 4, pp. 557–558, Feb. 2009.
[10] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J. D. Fekete, and D. W. Fellner, "Visual analysis of large graphs: state-of-the-art and future research challenges," *Computer Graphics Forum*, vol. 30, no. 6, pp. 1719–1749, Sep. 2011.
[11] P. Mahe and J.-P. Vert, "Graph kernels based on tree patterns for molecules," *Machine Learning*, vol. 75, no. 1, pp. 3–35, Apr. 2009.
[12] G. W. Furnas, A fisheye follow-up: further reflections on focus context," in *Proc. SIGCHI Conference on HUMAN Factors in Computing Systems*. New York, NY, USA: ACM, 2006, pp. 999–1008.
[13] P. Pons and M. Latapy, "Post-processing hierarchical community structures: quality improvements and multi-scale view," *Theory Comput. Sci.*, vol. 412, no. 8-10, pp. 892–900, Mar. 2011.
[14] S. Mori and P. C. M. van Zijl, "Fiber tracking: principles and strategies, a technical review," *NMR Biomed.*, vol. 15, no. 7-8, pp. 468–480, 2002.
[15] S. Zhang, S. Correia, and D. H. Laidlaw, "Identifying white-matter fiber bundles in DTI data using an automated proximity-based fiber-clustering method," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 5, pp. 1044–1053, Sep. 2008.
[16] M. Cosottini, M. Giannelli, F. Vannozzi, I. Pesaresi, S. Piazza, Belmonte, and G. Siciliano, "Evaluation of corticospinal tract impairment in the brain of patients with amyotrophic lateral sclerosis by using diffusion tensor imaging acquisition schemes with different numbers of diffusion-weighting directions." *Journal of Computer assisted Tomography*, vol. 34, no. 5, pp. 746–750, 2010.
[17] C. A. Sage, W. Van Hecke, R. Peeters, J. Sijbers, W. Robberecht, Parizel, G. Marchal, A. Leemans, and S. Sunaert, "Quantitative diffusion tensor imaging in amyotrophic lateral sclerosis: revisited," *Human Brain Mapping*, vol. 30, no. 11, pp. 3657–3675, 2009.
[18] E. Olivetti, T. B. Nguyen, and P. Avesani, "Fast clustering for interactive tractography segmentation," in *Proc. the 3rd IEEE Intl Workshop on Pattern Recognition in NeuroImaging*, 2013, pp. 42-45.
[19] X. Wang, W. E. Grimson, and C.-F. F. Westin, "Tractography segmentation using a hierarchical Dirichlet processes mixture model," *NeuroImage*, vol. 54, no. 1, pp. 290–302, Jan 2011.
[20] E. Olivetti and P. Avesani, "Supervised segmentation of fiber tracts," in *Proc. 1st International Workshop, SIMBAD'11*, Venice, Italy, 2011, pp. 261–274.
[21] E. Garyfallidis, "Towards an accurate brain tractography," Ph.D. dissertation, University of Cambridge, 2012.
[22] E. Pekalska, P. Paclik, and R. P. W. Duin, "A generalized kernel approach to dissimilarity-based classification," *J. Mach. Learn. Res.*, vol. 2, pp. 175–211, 2002.
[23] E. Olivetti, T. B. Nguyen, and E. Garyfallidis, "The Approximation of the Dissimilarity Projection," in *Proc. IEEE Intl Workshop on Pattern Recognition in NeuroImaging*, pp. 85–88, 2012.
[24] E. Pekalska, R. Duin, and P. Paclik, "Prototype selection for dissimilarity-based classifiers," *Pattern Recognition*, vol. 39, no. 2, pp. 189–208, Feb. 2006.

Bao T. Nguyen received his MSc degree in Computer Vision from Ritsumeikan University, Japan. After six months working as a researcher at Foundation Bruno Kessler (FBK), Italy, he became a PhD candidate at Trento University, Italy. His research interests are machine learning, image processing, medical imaging and bioinformatics. He is currently working on the understanding of the brain, an inter-disciplinary project of computer science and cognitive science, at NiLab, a joint research unit between FBK and Trento University. He is also a member of ACM and the IEEE Computer Society.

Emanuele Olivetti received his master's degree in physics and his Ph.D. in computer science from the University of Trento, Italy. He is a researcher at the Bruno Kessler Foundation (FBK) working on machine learning for neuroimaging experiments jointly with the local center for mind and brain sciences (CIMeC) within the University of Trento. His research interests include brain decoding, learning algorithms for diffusion MRI data, joint analysis of multiple neuroimaging data sources, active learning and Bayesian inference.

Paolo Avesani received his Dr. Degree in Information Science from the University of Milan, Italy. He is a researcher at Fondazione Bruno Kessler (FBK), in Trento, Italy, where he is leading NILab, a Neuroinformatics Laboratory, a joint initiative of FBK and the Center for Mind/Brain Sciences of the University of Trento. His research interests include statistical learning for brain decoding, multivariate methods for longitudinal brain mapping and learning learning algorithms for for brain connectivity.