iVAR: Interactive Visual Analytics of Radiomics Features from Large-Scale Medical Images

Lina Yu, Hengle Jiang, Hongfeng Yu, Chi Zhang
University of Nebraska-Lincoln

Josiah Mcallister, Dandan Zheng
University of Nebraska Medical Center

***Abstract*—Medical imaging enables researchers and practitioner to uncover the characteristics of diseases (e.g., human cancer) in great detail. However, the sheer size of resulting imaging data and the high dimension of derived features become a major challenge in data analysis, diagnosis, and knowledge discovery. We present a novel visual analytics system, named iVAR, targeted at observing the comprehensive quantification of tumor phenotypes by effectively exploring a large number of quantitative image features. Our system is comprised of multiple linked views combining visualization of three-dimensional volumes and tumors reconstructed by computed tomography (CT) images, and a radiomic analysis of high-dimensional features quantifying tumor image intensity, shape and texture, and three non-image clinical features. Thus, it offers insights into the overall distribution of quantitative imaging features and also enables detailed analysis of the relationship between features. We demonstrate our system through use case scenarios on a real-world large-scale CT dataset with lung cancer.**

***Index Terms*—visual analytics, interactivity, medical images, radiomics, high-dimensional data**

I. Introduction

Medical imaging, a noninvasive and visual means to access the characteristics of human tissues, is routinely applied in clinical practice for oncologic diagnosis and treatment guidance. Conventionally, tumor response to therapy is mostly measured using two-dimensional descriptors of tumor size. Although a change in tumor size is an important indicator depicting tumor phenotypic dynamics, other features, such as the intensity, shape, and texture of tumor, can also be significant. To this end, researchers have proposed *radiomics* [1], [2] to improve the quantification of tumor phenotypes noninvasively by applying a large number of quantitative features derived from medical images. However, when analysts attempt to analyze and understand medical imaging data, the sheer size of data and the high dimension of derived radiomics features become a major challenge: Tracking multiple patients over time using high-resolution medical imaging can generate tera- or even peta-scale data, and possibly result in hundreds or thousands of derived radiomics features. It is a non-trivial task for analysts to tackle large-scale medical imaging data and find the important features.

Although using automated approaches such as correlation clustering [3] and subspace clustering [4] is a viable alternative, it is hard for analysts to control these automated analytical processes. In addition, these clustering results often target a specific pattern while analysts may be interested in much more, such as correlated and uncorrelated dimensions, outliers, and so on. Once the interesting patterns are found in the subspace, it is necessary for analysts to track back to the details of the patients' imaging data to verify the results. Thus, analysts often desire to visualize the overall distribution of features and explore the relationship between paired or multiple features in an interactive way, especially, to possibly correlate imaging-derived features and clinical features.

We propose an interactive visual analytics system, named *iVAR*, to tackle large-scale medical imaging data and high-dimensional radiomics features. Our system can effectively and efficiently visualize three-dimensional (3D) volume data and organ objects reconstructed by computed tomography (CT) images, as well as for exploring the comprehensive quantification of tumor phenotypes using a large number of imaging-derived features. The major contributions of our interactive visual analytics framework are:

- We reconstruct 3D volume and tumor objects from medical images based on doctors' depictions. Our 3D interactive visualization tool brings intuitive insights (e.g., location and size) of a tumor to doctors.
- We derive compressive imaging features based on the first order statistics, shape, size, and texture of the 3D reconstruction of tumor objects, which provide a detailed characterization of tumor phenotypes.
- Our system supports an overview of the distribution of features, allows analysts to build their own subspace of custom paired features out of the large dimensional data space, and enables multiple features analysis tasks through the interaction with these features.

We present a use case of analyzing radiomic features derived from a CT dataset of patients with lung cancer. It demonstrates that our iVAR system can help users effectively explore a high dimensional feature space and identify features of interest, thereby facilitating the involvement of users in the analysis process and leading to possible new discoveries.

II. Related Work

There has been a great amount of research devoted to the visualization and analysis of medical information. Our work focuses on user interactions with imaging-derived features and the guidance of reducing a large number of dimensions into fewer features to explore the relationship within multi-dimensional views of interest.

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

A. CT-based Radiomic Signatures

It is possible to capture tumor phenotypic characteristics non-invasively, given advanced technologies in medical imaging. CT is one of the most widely used imaging modalities. The quantitative characteristics of tissue density makes CT imaging routinely used in cancer diagnoses and treatments. Tumor phenotypic differences (e.g., shapes irregularity, infiltration, heterogeneity, or necrosis) can be quantified in CT images using radiomic features. Radiomics [1], [2] provides a comprehensive quantification of tumor phenotypes by analyzing a large set of quantitative data characterization algorithms. Several studies [5], [6], [7], [8] have demonstrated that radiomics has a significant clinical potential to augment diagnosis, as well as improve tumor staging and therapy response assessment in oncological practice using medical images that can be routinely acquired at low costs.

B. Visual Analysis of Physical and Feature Spaces

Many statistical and visualization methods have been developed to facilitate scientists to study the correspondence of structures and behaviors. For example, WEAVE [9] combined the visualization of simulation data, measurement data, and 3D anatomical data concerning the propagation of excitation in the heart. Raidou et al. [10] proposed a visual tool to explore and visualize the feature space of imaging-derived tissue characteristics. The tool can also identify, explore, and analyze heterogeneous intra-tumor regions. In our work, we use reproducible imaging features from the field of radiomics [1] [11]. These features comprehensively represent quantitative information about intensity, shape, size, and texture of imaging data.

C. High Dimensional Data Analysis

Different approaches have been used to handle the curse of dimensionality problem in high dimensional data analysis. One solution is to use dimensionality reduction methods, such as principal component analysis (PCA) [12], random forest [13], t-SNE [14], and so on. In practice, domain experts need to use their domain knowledge to understand which combinations of dimensions make sense. They desire a representation of data with minimal information loss. For visual approaches to dimensionality reduction, many techniques are proposed. Johansson et al. [15] introduced a dimensionality reduction system based on user-defined quality metrics using weight functions to preserve as many important structures as possible. Turkay et al. [16] presented dual analysis model to allow the interactions in items and dimensions spaces. It enabled the joint interactive visual analysis of multivariate datasets with respect to their dimensions and actual data values.

III. System Design

We develop an interactive visual analytics system iVAR for exploration of 3D medical data and associated medical-imaging derived features.

TABLE I
Detailed Derived Features Grouped by Three Feature Types

First Order Statistics	Skewness, Uniformity, MeanAbsoluteDeviation, Energy, RobustMeanAbsolute, Deviation, Median, TotalEnergy, Maximum, RootMeanSquared, 90Percentile, Minimum, Entropy, StandardDeviation, Range, Variance, 10Percentile, Kurtosis, Mean
Shape and Size	Maximum3DDiameter, Compactness2, Maximum2DDiameter Slice, Sphericity, MinorAxis, Compactness1, Elongation, SurfaceVolumeRatio, Volume, SphericalDisproportion, MajorAxis, LeastAxis, Flatness, SurfaceArea, Maximum2DDiameter Column, Maximum2DDiameter Row, InterquartileRange
Textural Features	ShortRunLowGrayLevel Emphasis, GrayLevelVariance, LowGrayLevelRun Emphasis, GrayLevelNonUniformity Normalized, RunVariance, GrayLevelNonUniformity, LongRunEmphasis, ShortRunHighGrayLevel Emphasis, RunLengthNonUniformity, ShortRunEmphasis, LongRunHighGrayLevel Emphasis, RunPercentage, LongRunLowGrayLevel Emphasis, RunEntropy, HighGrayLevelRun Emphasis, RunLengthNonUniformity Normalized, GrayLevelVariance, SmallAreaHighGrayLevel Emphasis, GrayLevelNonUniformity Normalized, SizeZoneNonUniformity Normalized, SizeZoneNonUniformity, GrayLevelNonUniformity, LargeAreaEmphasis, ZoneVariance, ZonePercentage, LargeAreaLowGrayLevel Emphasis, LargeAreaHighGrayLevel Emphasis, HighGrayLevelZone Emphasis, SmallAreaEmphasis, LowGrayLevelZone Emphasis, ZoneEntropy, SmallAreaLowGrayLevel Emphasis, SumVariance, Homogeneity1, Homogeneity2, ClusterShade, MaximumProbability, Idmn, Contrast, DifferenceEntropy, InverseVariance, Dissimilarity, SumAverage, DifferenceVariance, Idn, Idm, Correlation, Autocorrelation, SumEntropy, AverageIntensity, Energy, SumSquares, ClusterProminence, Entropy, Imc2, Imc1, DifferenceAverage, Id, ClusterTendency

A. Input Data

We use computed tomography(CT) data of patients with lung cancer [17] in our current system development. Each CT scan data consists of a set of axial images with the resolution of 512×512 and the size of data for each patient can be up to 453 MB. The dataset also contains clinical information such as gender, age, survival time, and overall stage.

B. Radiomic Features

We adopt and extend the radiomic features defined in the existing work [5] to describe tumor phenotype characteristics. With our application and other potential domains in mind, we use 94 features as shown in Table I and group these features into the following three categories:

- *First order statistics.* This group of features describes the distribution of tumor image intensities. It consists of the computation of energy, entropy, kurtosis, skewness, and so on. These features help us know the histogram dispersion, asymmetry, and sharpness of tumor intensities.
- *Shape and size.* We include the descriptors of the 3D structure of a tumor region and determine the shape and size based features, for example, maximum 3D diameter, surface area, volume (provide lesion size information), compactness, spherical disproportion, sphericity, and surface to volume ratio (describe how spherical, or elongated the shape of the tumor is).
- *Textural features.* The first two feature groups do not provide any information regarding the relative position of the various gray-levels over an image. Texture features are based primarily on gray level co-occurrence (GLCM) and gray level run-length (GLRLM) texture matrices. Such features describe patterns or the spatial distribution of voxel intensities.

C. Visual Analytics Requirements

Based on the information provided by our collaborated domain experts, we have identified the following application requirements:

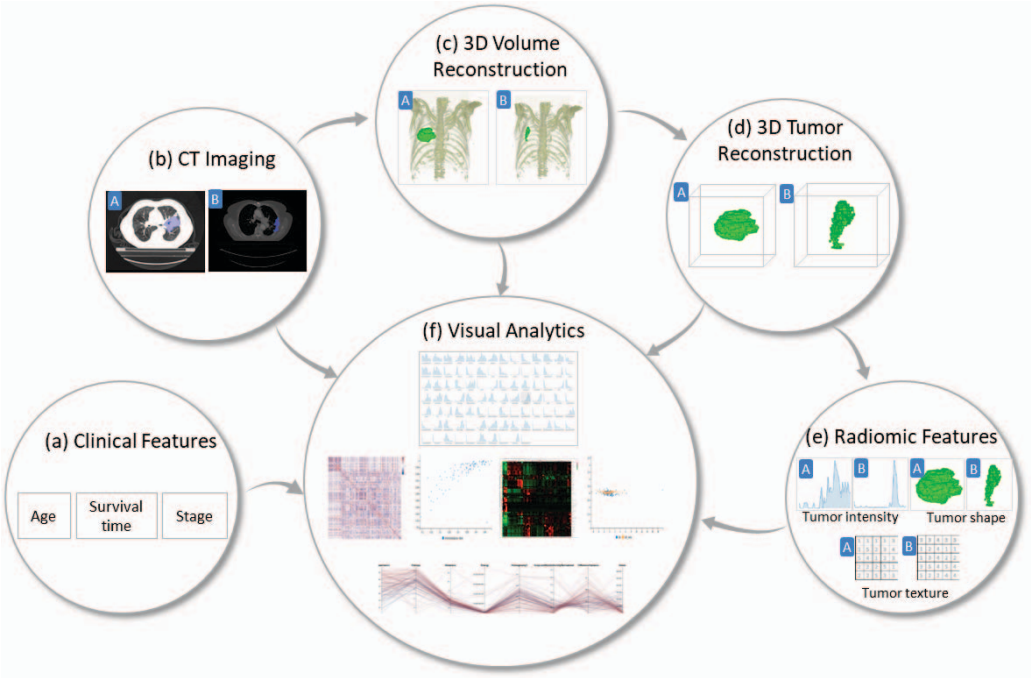

Fig. 1. An overview of the key steps of our visual analytics workflow using medical images, extracted radiomic features, and clinical features. (a) Non-image features are collected from clinical data. (b) Experienced physicians contour tumor areas on 2D CT slices, given two examples cases (A and B) in lung cancer patients. 3D volume (c) and tumor objects (d) are extracted according to 2D contours. (e) Radiomic features are extracted from the reconstructed tumors, quantifying tumor intensity, shape, size, and texture. (f) Linked visualization tools are used in the analysis of the features. Users can track back to any of the key stages (a)-(e) to interact with the data from clinical non-imaging features to 3D constructions to imaging-derived radiomic features.

R1 *Three-dimensional visualization and comparison.* Given a CT imaging dataset, a visualization of the 3D volume and the tumor can help in anatomical understanding for diagnosis, radiation therapy and surgical planning, and can also be used for education purposes. The 3D visualization methods also provide intuitive comparisons of shape and location of tumors among different patients.

R2 *Dimensionality reduction for derived features.* It is challenging for users to effectively study a large number of imaging-derived features. Automated dimensionality reduction methods, such as clustering, suffer from a lack of transparency and interpretability for domain experts who are not trained in statistics or machine learning. A visual exploration tool is desired to help them identify interesting patterns for building and refining multidimensional data.

R3 *Outlier detection.* The identification of anomalous data helps users not only detect the accuracy of data collection, but also identify the useless features.

R4 *User-driven feature relation exploration.* Multivariate exploratory visualization techniques are necessary to help users discover the relations among interesting combinations of dimensions.

D. Visual Analytics Design

1) Overview: We develop our visual analytics framework according to the identified application requirements as shown in Figure 1.

The first step is to collect non-image features from the clinical data shown in Figure 1(a) and 2D imaging datasets from CT scans shown in Figure 1(b). Experienced physicians contour the tumor areas on 2D CT slices.

In the second step, for each patient's imaging data, we reconstruct the 3D volume by inserting CT slices in sequence and do interpolation according to the thickness between two consecutive. The contour lines depicted by physicians in slices help us reconstruct the tumor volume by only filling the values within the contour lines. A direct volume rendering method, also called volume ray casting [18], is used for visualizing the whole volume shown in Figure 1(c) and the tumor objects shown in Figure 1(d). The rays are cast through the object and the 3D scalar fields of interest are sampled along the rays inside the object, so that we can render 3D scenes to 2D images. Our multivariate visualization technique can display the whole volume and highlight the tumor simultaneously. Thus, the 3D volume and the tumor object reconstruction can assist domain experts in getting an intuitive visualization of the shape and the location of the tumor in a patient's body. Our tool also supports the comparison among different cases. When domain experts can directly visualize the high-dimension, large-quantity data, erroneous intermediate results can be more easily detected to avoid the most challenging "garbage-in, garbage-out" problem facing big-data studies.

In addition, we derive 94 CT imaging features in the third step as described in Section III-B after extracting the tumor information, as shown in Figure 1(e). Here, we show two

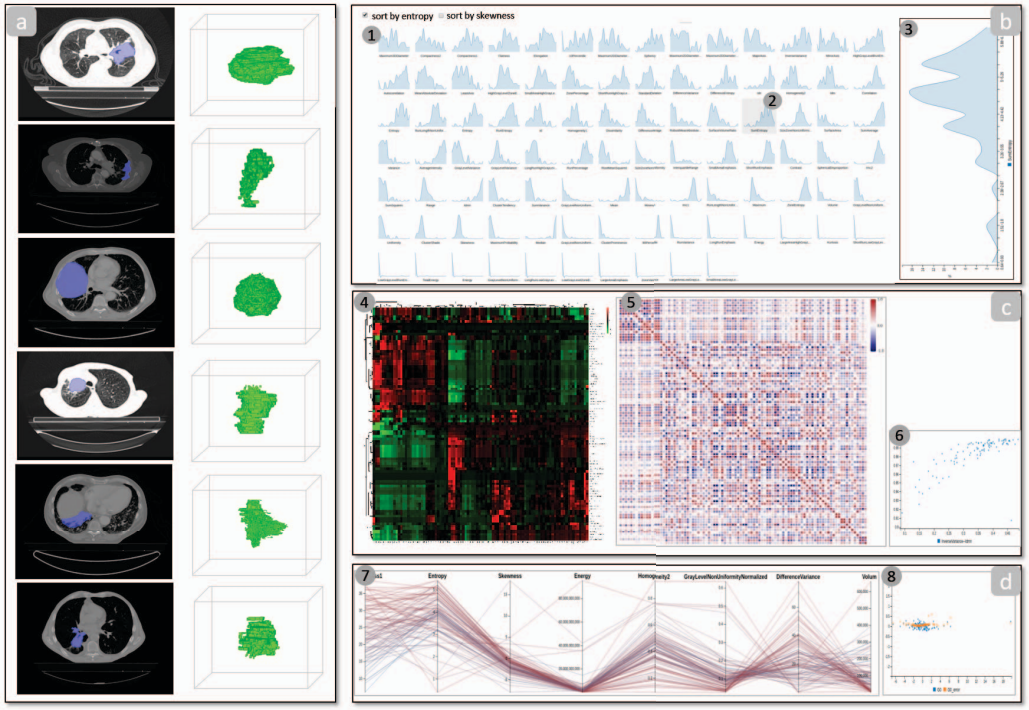

Fig. 2. The user interface of feature analytics. In the left panel, the view (a) displays the visualization of 2D CT slices and 3D tumor reconstructions. In the view (b), the panel (1) shows all dimensions as small multiple frequency plots. Selected dimensions are highlighted (2) and shown in a magnified plot (3). In the view (c), a heat map (4) and a correlation matrix (5) are introduced for dimensionality reduction and outliers detection. A scatter plot (6) shows the correlation between any two selected features. In the view (d), the multivariate panel consists of parallel coordinates (7) and a PCA scatter plot (8).

representative patients' cases as examples.

In the fourth step, we perform interactive visual analytics of different features. Apart from the quantitative image features, the original CT images, and the 3D constructions, clinical non-image features (as shown in Figure 1(a)) are also included to convey the relationships between clinical data and imaging data. As shown in Figure 1(f), we apply several visualization tools and analysis methods to the analysis of features to meet the requirements proposed by our collaborated domain experts in Section III-C. We describe these visualization tools in details in Section III-D2. In this step, users can track back to any available data, such as 3D volume and tumor reconstructions, to verify their observations.

2) Visual Representations and User Interface: We implement our framework through a number of highly interactive linked views shown in Figure 2. The layout of the user interface is broken down into two control panels: the left panel in Figure 2 is the 3D visualization panel, and the right three views constitute the feature analysis panel. Once a user selects a number of subjects loaded to the system, the tumor information is computed according to the contours depicted by physicians. The selected 3D rendering results are shown in the 3D visualization view (Figure 2(a)) so that the user can directly gain the 3D geometry information (e.g., location and size) of the tumors.

The view of Figure 2(b) shows the plot of each feature's frequency across all the patients, which allows analysts to quickly grasp the distributions of features. Typically, uniform distributions represent high uncertainty or high entropy while non-uniform distributions denote low uncertainty. In the case of non-uniform distributions, it is useful for analysts to know whether there is a positive or negative skew, which shows whether high or low values dominate. Therefore, we sort small multiple frequency plots in a descending order by the values of entropy or skewness. On the right of the small multiple frequency plots, a magnified frequency plot (3) shows the currently selected dimension (2). The mixed attribute types, such as numerical, ordinal, categorical data, can be flexibly represented here. We also provide different scales (i.e., linear or logarithmic) for the representation of the data values.

To identify the outliers and further reduce the dimensions, we introduce a radiomics heat map in the panel (4) with hierarchical clustering results in the view of Figure 2(c). In the heat map, the horizontal axis represents the patients, and the vertical axis represents the 94 radiomics features. We compute each entry of the heat map that corresponds to the value of a feature of a patient. For a feature f, we first normalize all its values across all the patients. Then, we compute the mean avg_f and the standard deviation $stddev_f$ of f. For a given patient i, we compute its value v_{f_i} as $(f_i - avg_f)/stddev_f$, where f_i is the original feature f value of the patient i. Here, we assume that the feature f has a normal distribution for all patients. If a normalized value v_{f_i} is larger than 3, then we consider it as an outlier. We know that for a normal distribution about 99.7% points lie in $[avg_f - 3stddv_f, avg_f + 3stddv_f]$. For each value v_{f_i} of a feature f of a patient i, we map

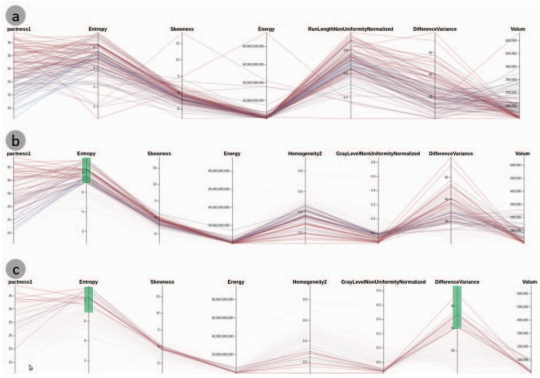

Fig. 3. Exploration of multidimensional patterns by parallel coordinates.

it linearly to the color from green ($\leq avg_f - 3stddv_f$) to black ($= 0$) to red ($\geq avg_f + 3stddv_f$). Thus, outliers are 0.3% that correspond to either green or red and can be easily identified from the heat map. In order to further identify the patterns across the patients and the features, we apply unsupervised hierarchical clustering along the horizontal and vertical axes. The resulting view shows the clusters of patients with similar radiomic patterns. Therefore, it helps users to find their interested patterns and important features.

To deal with the curse of dimensionality, we also show a correlation matrix panel (5) in the view of Figure 2(c). Users can easily investigate the dependence between multiple features, so that the useless and the most significant dimensions can be easily discovered. Each entry in the matrix expresses the Pearson correlation coefficient of the corresponding pair of features. Users can click on any block in the correlation matrix, a scatter plot will be displayed in the panel (6) to show the relationship between two corresponding features selected from the correlation matrix.

Once users find their interested features or obtained the reduced dimensions, a parallel coordinates panel (7) in the view of Figure 2(d) can be used to show the feature values in multiple dimensional spaces. Parallel coordinates plot patients' data across many feature dimensions. Each of the feature dimensions corresponds to a vertical axis and each patient's data is displayed as a polyline along the dimensions. In this view, features can be quickly compared by filtering along any dimension. Users can click and drag along a given dimension to update the filter. Figure 3 shows the interactive results of the exploration of multidimensional patterns by parallel coordinates. Figure 3(a) loads all patients data from seven feature dimensions. When a user brushes the values larger than 7 on the `Entropy` dimension, Figure 3(b) shows the corresponding values on other dimensions. When the user continues to brush the high values on the `DifferenceVariance` dimension, Figure 3(c) shows the further results.

Parallel coordinates may have a few issues in certain circumstances. For example, meaningful patterns can be obscured by a clutter of lines, especially with large datasets and the order of the axes, which can impact how users understand the data. To address these issues, we introduce the principal component analysis (PCA) panel (8) in the view of Figure 2(d) to emphasize variation and bring out strong patterns in user selected features.

All the panels are performed in a linked fashion so that when users select or brush data elements in a panel, the other panels will update their plots of the corresponding data elements. In this way, users can intuitively interact with their analysis processes and interpret results.

IV. Application and Results

The Lung1 dataset [17] consists of 422 non-small cell lung cancer (NSCLC) patients that were treated at Department of Radiation Oncology (MAASTRO) Clinic, The Netherlands. CT scans, manual delineations, clinical data (e.g., survival time) are available in this dataset. In this section, we use the data of 100 patients from this dataset to show a usage scenario of our visual analytics system, and demonstrate the usefulness of our design by a set of highly interactive linked views to explore the relationship among radiomic features.

A. Feature Overview

The analyst selects multiple subjects from the database, the corresponding multiple 3D volume data and the tumor objects data will be computed by the original imaging data. Meanwhile, the computed feature distributions will be automatically populated in the feature analysis panel. Small multiple frequency plots are sorted in a descending order by the values of entropy. According to the order of features' frequency, the analyst can easily find the features with rich information and with a small amount of information, which assists the analyst in eliminating the useless features in building the subspace.

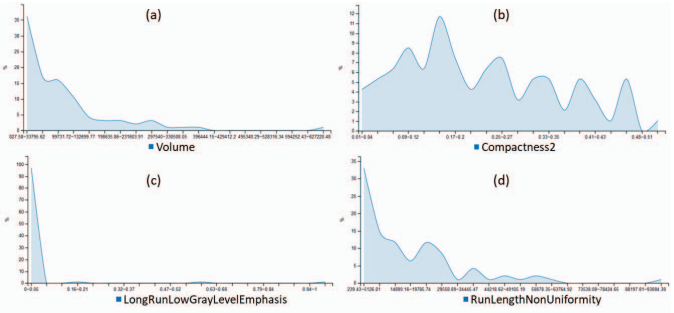

Fig. 4. The frequency plots of four selected features: (a) `Volume`, (b) `Compactness2`, (c) `LongRunLowGrayLevelEmphasis`, and (d) `RunLengthNonUniformity`.

For example, Figure 4(a) shows the plot of the feature `Volume`, a commonly used measurement of tumor volume, with a certain variation among the patients. Figure 4(b) shows the plot of `Compactness2` that is a derived radiomic feature quantifying how compact the tumor shape is. Although `Volume` and `Compactness2` belong to the same feature category *shape and size* (see Section III-B), we can clearly see that `Compactness2` exhibits more information in its plot and can be used as a more effective radiomic signature.

Similarly, from the overview of the frequency plots, we can easily observe that the plots of several features in the

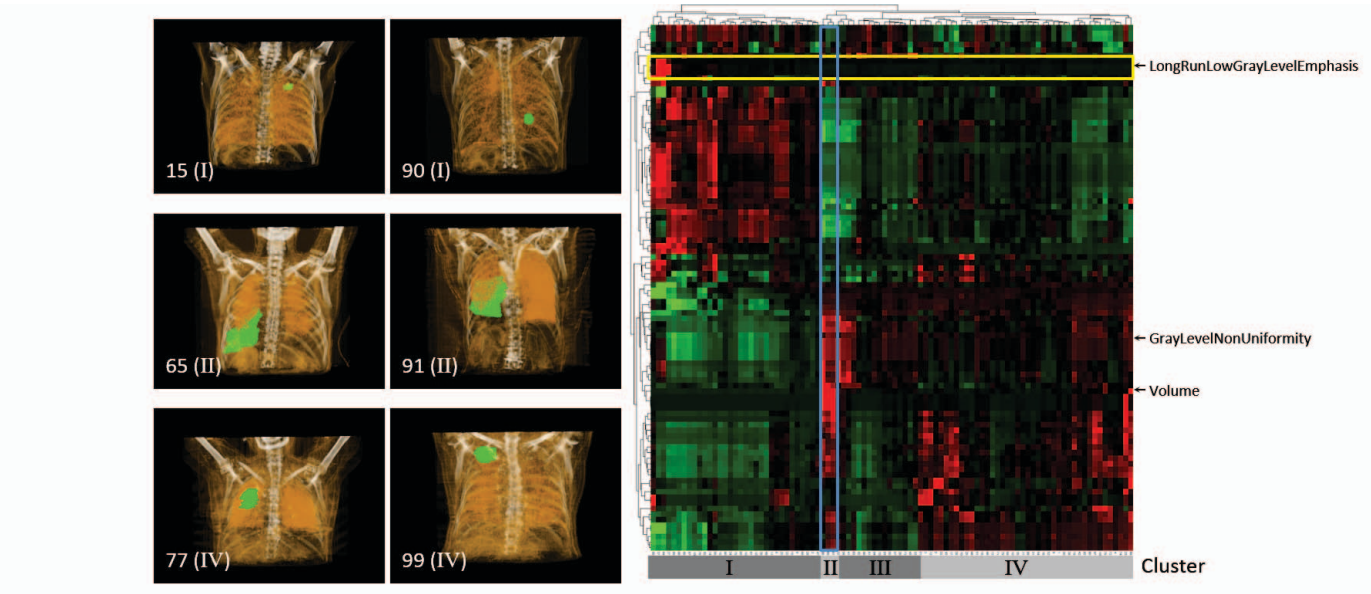

Fig. 5. Radiomic heat map with unsupervised hierarchical clustering of the lung cancer patients along the horizontal axis and the radiomic features along the vertical axis, revealing the clusters of patients with similar radiomic expression patterns. The 3D volume visualization can reveal the detailed 3D geometrical structures of tumors in different clusters. In each image, the first number is the patient index, followed by the cluster index in parentheses.

category *textural features* convey a small amount of information. For example, Figure 4(c) shows the plot of `LongRunLowGrayLevelEmphasis` that is a feature indicating whether the tumor texture is dominated by long runs with low gray levels or not. Figure 4(d) shows the plot of *RunLengthNonUniformity* that is a feature in the same category and measures the intratumor heterogeneity. We can clearly see that *RunLengthNonUniformity* conveys a significantly higher amount of information and can be a candidate of radiomic signature.

B. Feature Pattern

To investigate detailed feature patterns, the analyst can explore the radiomic heat map. It shows not only the values of patients across all features, but also the cluster information on both patients and features. This map can facilitate the analyst to identify feature patterns and develop her understanding of features. In this case, we show the heat map in Figure 5. It can be observed that the features shown in the yellow box are clustered in the same cluster, most of patients' values in these features are shown in black. We can easily infer that these features have very stable values and low entropy on their distributions. These features, including `LongRunLowGrayLevelEmphasis`, existentially belong to the category *textural features*, and their distributions can be easily verified in the overview of the frequency plots.

The heat map in Figure 5 approximately conveys four clusters of patients according to the similarities of their values of all the features. For each cluster, the analyst can track back to the details of the patients' imaging data and the 3D reconstructions. Figure 5 shows the volume visualization of a few selected patients in different clusters. We can clearly see that the tumors convey similar structural characteristics within the same cluster, but noticeable differences across the clusters. For example, the heat map successfully captures the significantly distinct phenotypes of the tumors in Cluster II (i.e., the green and red feature values in the blue box)(see Section III-D2), which is intuitively verified using the volume visualization, as shown in the plots of the patients #65 and #91 in Figure 5.

C. Feature Correlation

After gaining the patterns of the features, the analyst can choose the features in the clusters and further explore the relationship between the features. In particular, identifying the association of radiomic expression patterns with tumor stage and survival time is essential for capturing prognostic radiomic signatures and developing predictive models of survival.

The correlation matrix in our system, displaying the correlation coefficient of each pair of features, can facilitate the analyst to first quickly grasp an overview of the relationship among the features and then examine the details. Figure 6 shows the correlation matrix of the clinical features and the radiomic features. The clinical features include `age`, `gender`, `T-stage`, `N-stage`, `Overall stage`, and `Survival time`. From the correlation matrix, we can clearly see that there are strong correlations among the most shape and size features. The only exceptions are `Elongation` and `Flatness`, which are strongly correlated with each other but not other features. These two features also do not convey a significant correlation with `Survival time`. Therefore, the analyst can possibly neglect them when selecting radiomic signatures. Meanwhile, we can examine the column of `Survival time` and identify the features that are considerably associated with `Survival time` (e.g., `Volume`). By clicking each entry in the matrix, our system

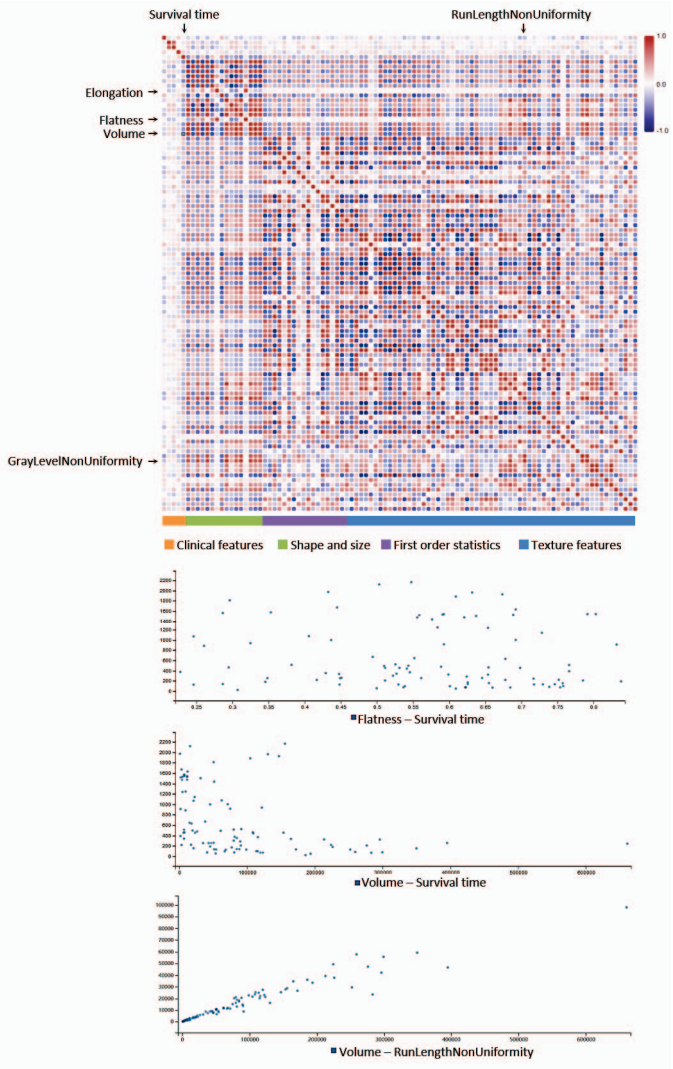

Fig. 6. The correlation matrix containing the clinical features, the shape and size features, the first order statistics features, and the texture features (top), and the two scatter plots of the select pairs of features (bottom).

can display the scatter plot of the corresponding pair of features. Figure 6 shows the example plots of `Survival time` and `Flatness`, `Survival time` and `Volume`, and `Volume` and `RunLengthNonUniformity`, respectively. These plots match the corresponding correlation coefficients in the matrix, and can help us gain a more detailed view of the relationship between two features. For example, the scatter plot of `Volume` and `RunLengthNonUniformity` shows a strong positive correlation, which matches the red color at the corresponding entry in the matrix. In this way, the analyst can effectively analyze any pair of features, and identify potential associations.

D. Feature Subspace

Based on the feature pattern and correlation analysis, our analyst can then explore suggested subspaces and decide which one to focus. Our system provides parallel coordinates

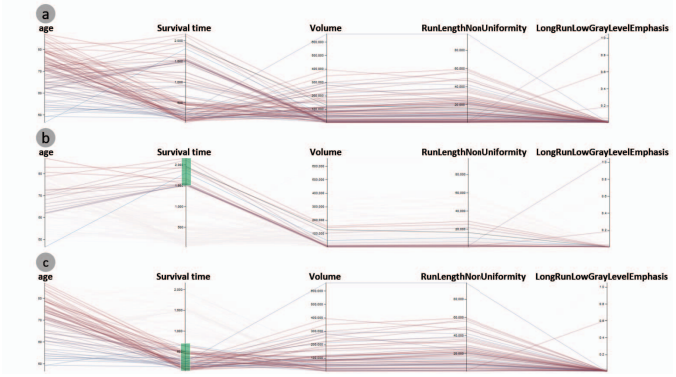

Fig. 7. The parallel coordinates plot of five features (a), and the brushing results (b and c).

and principal component analysis (PCA) with brushing to assist the analyst in studying feature subspaces.

For illustration purposes, Figure 7(a) shows a parallel coordinates plot of a feature subspace consisting of `age`, `Survival time`, `Volume`, `RunLengthNonUniformity`, and `LongRunLowGrayLevelEmphasis`. From the overall of frequency plots (see Figure 4), we have already known that the feature `LongRunLowGrayLevelEmphasis` contains a less amount of information and thus may be less of interest. This can be clearly perceived in Figure 7(a) with a narrow range of the `LongRunLowGrayLevelEmphasis` values for most patients. Similarly, there is no strong correlation between `age` and `Survival time`, but a clearly linear correlation between `Volume` and `RunLengthNonUniformity`, which matches the scatter plot in Figure 6.

However, Figure 6 reveals a negative correlation between `Survival time` and `Volume`, which is not obvious in Figure 7(a). By brushing the `Survival time` axis, we can closely examine the association with `Survival time`. In Figure 7(b), the analyst brushes the high survival time, and the resulting polylines clearly show a negative correlation between `Survival time` and `Volume`. When the analyst interactively moves the brushed area towards lower `Survival time` values, the correlation declines, as shown in Figure 7(c). Therefore, the parallel coordinates plot allows users to interactively gain a deeper understanding of the relationships among the features.

The analyst can also use the PCA tool of our system to develop predictive models between features. For example, the heat map in Figure 5 shows that the texture feature `GrayLevelNonUniformity` and the shape and size feature `Volume` are similar. In addition, Figure 6 shows that these two features are correlated with `Survival time`. By applying PCA, the analyst can obtain the result conveying the nonlinear relationship between `GrayLevelNonUniformity`, `Volume`, and `Survival time`, as shown in Figure 8. This technique can be also applied to other features of interest.

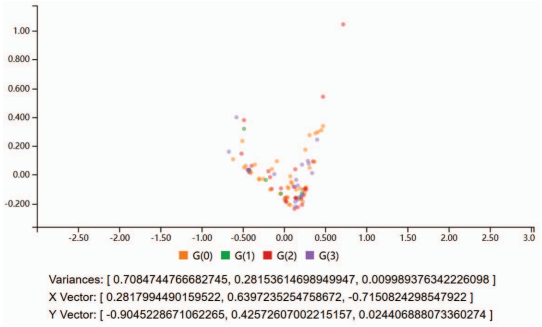

Fig. 8. The PCA result of `Survival time`, `Volume`, and `GrayLevelNonUniformity`. The variances and the two orthogonal basis vectors are also provided. The data points can be colored according to other features. In this example, the colors correspond to four N-stages.

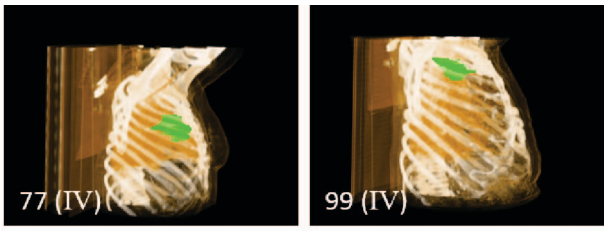

Fig. 9. The side view of the 3D volume visualization.

V. Conclusion

We propose a visual analytics system, named iVAR, for domain experts to interactively build and refine high-dimensional features derived from large medical imaging data with the help of the suggestion from the system. As shown in the existing work [2], [5], automated approaches can help analysts mine essential information from a large number of features. In this work, we demonstrate that finer-grain information can be further captured through interactive visual analytics. Through our iVAR system with a linked view design, users can interactively and simultaneously investigate multiple aspects of a large feature space, and possibly obtain new insights that could not be extracted using automated approaches.

In the future, we plan to exploit more high dimensional data visualization techniques to enhance the capability of our iVAR system. Moreover, in our current study, we note that the existing radiomic approaches [2], [5] mostly employ the features derived from the *local* information of tumors, and cannot capture *global* structure information. For example, as shown in Figure 5, the radiomic features of the patients #77 and #99 exhibit a significant similarity, and are tightly coupled to one cluster. However, the survival times of these two patients are 58 days and 493 days, respectively, showing a less optimal performance of the derived radiomic features. By closely examining their 3D structures (Figure 9), we observe that these two tumors have different global structural characteristics (e.g., location), although their local characteristics are very similar. We would like to investigate new global information based approaches [19] to improving the prognostic performance of radiomic signatures.

Acknowledgment

This research has been sponsored by the University of Nebraska Medical Center (UNMC) Faculty Diversity Fund.

References

[1] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher *et al.*, "Radiomics: the process and the challenges," *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[2] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker *et al.*, "Radiomics: extracting more information from medical images using advanced feature analysis," *European journal of cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[3] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Machine Learning*, vol. 56, no. 1-3, pp. 89–113, 2004.

[4] L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 90–105, 2004.

[5] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature communications*, vol. 5.

[6] F. Davnall, C. S. Yip, G. Ljungqvist, M. Selmi, F. Ng, B. Sanghera, B. Ganeshan, K. A. Miles, G. J. Cook, and V. Goh, "Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice?" *Insights into imaging*, vol. 3, no. 6, pp. 573–589, 2012.

[7] Y. Balagurunathan, Y. Gu, H. Wang, V. Kumar, O. Grove, S. Hawkins, J. Kim, D. B. Goldgof, L. O. Hall, R. A. Gatenby *et al.*, "Reproducibility and prognosis of quantitative features extracted from CT images," *Translational oncology*, vol. 7, no. 1, pp. 72–87, 2014.

[8] D. V. Fried, S. L. Tucker, S. Zhou, Z. Liao, O. Mawlawi, G. Ibbott, and L. E. Court, "Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer," *International Journal of Radiation Oncology* Biology* Physics*, vol. 90, no. 4, pp. 834–842, 2014.

[9] D. L. Gresh, B. E. Rogowitz, R. L. Winslow, D. F. Scollan, and C. K. Yung, "WEAVE: A system for visually linking 3-D and statistical visualizations, applied to cardiac simulation and measurement data," in *Proceedings of the conference on Visualization'00.* IEEE Computer Society Press, 2000, pp. 489–492.

[10] R. G. Raidou, U. A. van der Heide, C. V. Dinh, G. Ghobadi, J. F. Kallehauge, M. Breeuwer, and A. Vilanova, "Visual analytics for the exploration of tumor tissue characterization," in *Computer Graphics Forum*, vol. 34, no. 3. Wiley Online Library, 2015, pp. 11–20.

[11] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2015.

[12] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.

[13] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[14] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[15] S. Johansson and J. Johansson, "Interactive dimensionality reduction through user-defined combinations of quality metrics," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 6, pp. 993–1000, 2009.

[16] C. Turkay, P. Filzmoser, and H. Hauser, "Brushing dimensions—A dual visual analysis model for high-dimensional data," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2591–2599, 2011.

[17] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Cavalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld *et al.*, "Data From NSCLC-Radiomics. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2015.PF0M9REI," 2015.

[18] S. D. Roth, "Ray casting for modeling solids," *Computer graphics and image processing*, vol. 18, no. 2, pp. 109–144, 1982.

[19] L. Yu and H. Yu, "Boundary-structure-aware transfer functions for volume classification," in *SIGGRAPH ASIA 2017 Symposium on Visualization*, ser. SA '17, 2017.