This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

A Visual Analytics Approach using the Exploration of Multi-Dimensional Feature Spaces for Content-based Medical Image Retrieval

Ashnil Kumar, *Member, IEEE,* Falk Nette, Karsten Klein, Michael Fulham, and Jinman Kim, *Member, IEEE.*

***Abstract*—Content-based image retrieval (CBIR) is a search technique based on the similarity of visual features and has demonstrated potential benefits for medical diagnosis, education, and research. However, clinical adoption of CBIR is partially hindered by the difference between the computed image similarity and the user's search intent, the semantic gap, with the end result that relevant images with outlier features may not be retrieved. Furthermore, most CBIR algorithms do not provide intuitive explanations as to why the retrieved images were considered similar to the query (e.g., which subset of features were similar), hence it is difficult for users to verify if relevant images, with a small subset of outlier features, were missed. Users therefore resort to examining irrelevant images and there are limited opportunities to discover these 'missed' images. In this paper, we propose a new approach to medical CBIR by enabling a guided visual exploration of the search space through a tool, called Visual Analytics for Medical Image Retrieval (VAMIR). The visual analytics approach facilitates interactive exploration of the entire dataset using the query image as a point-of-reference. We conducted a user study and several case studies to demonstrate the capabilities of VAMIR in the retrieval of CT images and multi-modality PET-CT images.**

***Index Terms*—content-based image retrieval, visual analytics, medical imaging, PET-CT**

I. Introduction

THE critical and indispensable role played by modern multi-dimensional (3D/4D) and multi-modal (e.g., PET-CT and PET-MR) medical images for patient diagnosis and disease monitoring has led to a rapid expansion in the volume of image data that are acquired and stored in clinical environments. The medical image databases in hospitals grow daily by several thousand images with consequent challenges for the storage, retrieval, and interpretation of these imaging data [1].

In recent years, one important focus of medical informatics research has been the retrieval of stored imaging data from these collections for use in clinical applications, education, and research [1]–[4]. Content-based image retrieval (CBIR) is an image search technique that can be used for such purposes. CBIR is characterised by the use of automatically extracted visual image features as the search criteria [5]. These image features include, but are not limited to, color, texture, shape, and the spatial arrangement of regions of interest (ROIs). In contrast to conventional text-based approaches, CBIR does not require manual annotation of the images, which is not feasible for the volume of medical images that are available in hospital databases today. One of the major limitations of CBIR is the semantic gap, i.e., the difference between the understanding and search intent of a human user and the computed (machine) understanding of the image content [6]. Hence, human interpretation of the retrieved data is an important component of the CBIR workflow. However, human interpretation of the retrieved images is still an emerging research area [1], [7], [8].

To overcome the semantic gap, several medical CBIR systems [9]–[15] retrieve the m most similar images (as decided by the underlying algorithm) and let the user be the final arbiter to determine which of the images are relevant to the query. Results are often presented to the user as a ranked list that is based on the computed similarity to the query. There is no guarantee that these ranked lists contain all the relevant images from the dataset. Even though most retrieval algorithms are focused on optimising precision, i.e., the proportion of retrieved images that are relevant, some relevant images may still be given a low ranking. Commonly used retrieval algorithms find between 5 to 20 images and so it is possible that in a large database a relevant image could fall below this threshold [12]. In routine practice, the interpretation of modern medical images is time-consuming given the volume of data, image resolutions, different acquisition parameters, and data from multiple modalities. Thus it is not feasible to simply expand the number of images included in a ranked list and users may not have the opportunity to consider low ranked images for their specific application. Ranked lists do not provide the user with an explanation about the features responsible for the rankings and do not explicitly communicate how the result set is distributed in the feature space or how well the list covers different types of similar images in the image database. For example, the algorithm may provide similar rankings for two images despite different characteristics, e.g., features representing tumor size versus number of tumors. It is important to know what criteria were used because the retrieved images might represent diverse subsets of the database. We hypothesise that a method to 'discover' images with low rankings may identify outlier cases that are still relevant to the query.

A. Kumar, F. Nette, K. Klein, M. Fulham, and J. Kim are with the School of Information Technologies, University of Sydney, Australia. E-mail: {ashnil.kumar, karsten.klein, michael.fulham, jinman.kim}@sydney.edu.au

K. Klein is now with Monash University, Australia.

F. Nette is also with Lübeck University, Germany.

M. Fulham is also with Sydney Medical School, University of Sydney, Australia and the Department of Molecular Imaging, Royal Prince Alfred Hospital, Sydney, Australia.

This work was supported in part by ARC grants.

K. Klein was also partly supported by Tom Sawyer Software and Newton-Green Technologies.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

Techniques that have been previously investigated to discover lowly ranked images include automatic feature selection, relevance feedback, and weighting and we outline these below. Automatic feature selection finds the most discriminative features using learning techniques from a labelled dataset [16], [17]. Generally, it is difficult to optimise feature selection for individual queries meaning that the feature space cannot be tuned for the natural diversity in different medical images. One technique for query-based feature selection is a hierarchical classification with the classifiers at different levels of the hierarchy having different sets of optimal features [18]. Dy et al. [18] used a two-phase classification approach: (i) the first phase categorised the query image into one of several classes, using features optimised for discriminating between classes and, (ii) the second phase categorised the query into one of several subclasses, using features optimised for discriminating between subclasses. In such an approach, a failed classification at a higher level, due to an image with outlier features, means that subsequent levels and the final retrieval would utilise suboptimal features.

An alternate way of performing query-based feature selection is through relevance feedback [13], [19], [20]. During relevance feedback a user marks a subset of the ranked list as being most relevant to the query. This subset is used to automatically select the most discriminative features in multiple iterations of the search. However, as with automatic feature selection, the user has no direct control over which features are employed and is inherently limited by the diversity of the images within the ranked list, i.e., a user cannot utilise the diversity of the entire dataset.

Finally, weighting is a process in which users manually assign different priorities to different features [14]. Weighting assumes that users have an in-depth understanding of how the image features relate to the retrieval algorithm, which is very unlikely. Users may not understand the outcome of weighting particular features and this may lead to unexpected retrieval results [21].

Lew et al. [22] suggested that the ability of humans to explore and understand retrieved images is hindered by the gap in search systems that leverage user experiences and knowledge. A technique for browsing retrieved images in the context of the image features would enable users to understand these images and allow them to refine the search based on their improved understanding of the characteristics used to define image similarity.

Visual analytics (VA) methods offer the opportunity to solve complex retrieval tasks by combining human and automated analyses. VA utilises interactive visualisations to facilitate analytical reasoning based upon information extracted from the database [23]. Hiroike et al. [24] proposed a VA technique where the high-dimensional image feature space was transformed to a 3D coordinate space to cluster thumbnails of images with similar features. Gao et al. [25] reported an approach that allows an interactive refinement of hypotheses for image classifier training in multimedia collections. The feature similarity of retrieved data was analysed by Rodrigues et al. [26], but this work used the entire feature space without the opportunity to disregard irrelevant features or to tune the feature space based on user knowledge.

VA and relevance feedback can be seen as complementary approaches to improve retrieval by bridging the semantic gap. It has been identified that the quality of retrieval can be improved through interfaces that allow users to: (i) interactively adjust the search parameters through a multimedia analytics approach, and (ii) provide relevance feedback on individual results [27]. Combining analytics and relevance feedback is not a trivial undertaking and it requires striking the right balance between real-time browsing and supporting the search through feedback loops [28].

In this paper, we present the integration of VA into a medical CBIR system, which we refer to as Visual Analytics for Medical Image Retrieval (VAMIR). We suggest that VAMIR will provide more effective and efficient retrieval and interpretation of modern medical images through a guided exploration of the large feature space using the query image as a point-of-reference. VAMIR will also enable the evaluation of 'missed' images from other techniques. In VAMIR, we have not combined VA with relevance feedback and we will consider it as a project for future research.

We envision that VAMIR will be primarily useful for scientific and education applications of medical CBIR, with potential for extension to clinical decision support, and as such our paper will focus on the value of VAMIR in these contexts. We suggest that VAMIR will be particularly useful for the analysis of volumetric (3D) and multi-modality images, where an individual patient scan consists of typically hundreds or thousands of 2D images. Thus we evaluated VAMIR on case studies and a user study on two datasets: the public collection of computed tomography (CT) volumes made available by the Lung Imaging Database Consortium and Image Database Resource Initiative (LIDC/IDRI) and a combined positron emission tomography and computed tomography (PET-CT) lung cancer database.

We previously reported a preliminary version of this work that described the abstract concept of VAMIR [29]. In this paper, we expand the preliminary data by providing the equations necessary for dynamically modifying our abstract visualisations, introducing automatic feature selection for the visualisations, enabling dynamic querying within the VAMIR interface, conducting a user study, and evaluating VAMIR with a different, larger dataset.

This paper is organised as follows. In Section II we describe our VA framework and how it can be integrated into the CBIR workflow. Section III lists the visualisation capabilities of VAMIR; in Section IV the interactive capabilities of VAMIR and how they can be used for exploring and retrieving images are outlined. We describe the datasets in Section V and implementations used for evaluation in Section VI. The procedures used to evaluate our work are described in Section VIII, with results in Section VIII and a discussion in Section IX. We summarise our contributions and list ideas for future investigation in Section X.

II. Visual Analytics Framework

The workflow of our VA framework for CBIR is shown in Fig. 1. The framework has a CBIR engine, visualisations for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

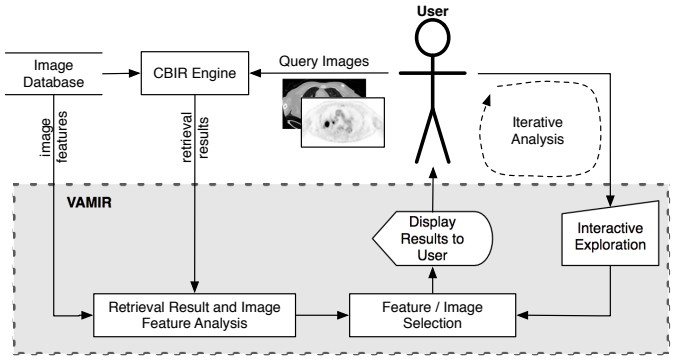

Fig. 1. The integration of the CBIR workflow and the VA concept. The user first carries out a query using a CBIR system. After retrieval, VAMIR enables the user to visually explore the search space. Users can iteratively refine the image feature space to find images that they consider relevant to the query.

the query image and the retrieved images, and VA components that allow the user to filter and select images and image features for exploration. A specific CBIR engine is not required in our workflow and different engines can be used for different datasets or applications.

The VAMIR components in the workflow form an interactive visualisation scheme for the analysis of retrieval results from the CBIR engine. Guidelines for visualisation [30] and search interfaces [31] suggest that effective human interpretation of visualised information can be enabled through multiple different views of complex data, abstractions to summarise complex information, the presentation of supplementary information, and the ability to interact with the visualisation to change the information that is presented or to refine the search space.

We developed several requirements for VAMIR based on these guidelines and our overall aim of using VA to explore the feature space. The requirements include:

1) providing multiple views of the feature space;
2) enabling automatic feature selection, based on user selection, to populate each of the views;
3) abstracting the individual images in each view to avoid rendering multiple large images;
4) providing the ability to examine images directly by linking each abstraction to its associated image;
5) allowing the user to refine the visualisation and,
6) enabling the user to refine or modify the search space and dynamically adjust the query.

A typical usage scenario for the workflow is as follows. A user inputs an image as a query to the CBIR engine. The information generated during the query process (similarity/dissimilarity of indexed images compared to the query image) and the features of the images from the search index are passed to VAMIR. The visualisation process with VAMIR allows the user to gain insights about the differences in the features of the query image and the images in the database. This enables users to find images by examining the entire dataset and to discover any images that may have been missed by the retrieval algorithm.

III. VAMIR Visualisation Design

The assumption underlying our visualisation design is that only a subset of the image features are important for a user to understand the retrieved images in relation to the query, i.e., a subset of the features has a higher priority for image similarity. The subset could change depending on the query, the user, or the final retrieval application. Thus, the VAMIR visualisation allows the user to see new patterns that are not apparent in a ranked list by grouping images based on their similarity using the feature subset.

In the VAMIR visualisation, we use nodes (points) as abstract representations of images. We group two nodes in close proximity if the difference between the subset of features is small. A collection of nodes in close proximity therefore indicates a group of images that share similar characteristics. Interaction with the visualisation allows the user to focus on regions of interest and to investigate the relation of distribution patterns across the linked feature groups as described below in Section IV. Users can then apply domain knowledge about the importance of image features, e.g., tumor size, to interactively adjust and refine the query or to explore different feature subsets.

VAMIR allows visualisation of the relationships between the query and retrieved images, and between the retrieved images themselves, based upon a selected feature subset. The overall similarity of the query and the retrieved images using all features is also displayed. The VAMIR interface is shown in Fig. 2; its functionality is described below.

A. Description

Our visualisation, shown in Fig. 3, is an abstract graphical representation of the image dataset within the feature space. The two major components of our visualisation are a central disc (hereafter referred to as the 'super-centre') and a set of 'sectors' surrounding the super-centre. The super-centre presents the overall similarity of the retrieved images compared to the query image as calculated by the CBIR engine. Each sector presents the image similarity for a subset of the features. Only two features are used as the subset for each sector to reduce the dimensionality of the feature space being visualised. We use different pairs of features, i.e., unique feature subsets, for each sector to ensure that the sectors provide complementary visualisations.

The polar coordinates of the nodes (the abstract representations of images) are determined according to the difference in the corresponding features in comparison to the query image. The user can examine the 2D spatial distribution of the nodes to assess the similarity of the features of images in the dataset compared to the features of the query image. In contrast, a projection of the entire feature vector in the plane, e.g. by multi-dimensional scaling (MDS) would obscure the similarity (or dissimilarity) of individual features.

The node representing the query image is placed on the outer arc of the super-centre to support easy orientation. Similarly, the node representing the query image is placed in a corner of the outer arc of every sector. All other image nodes

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

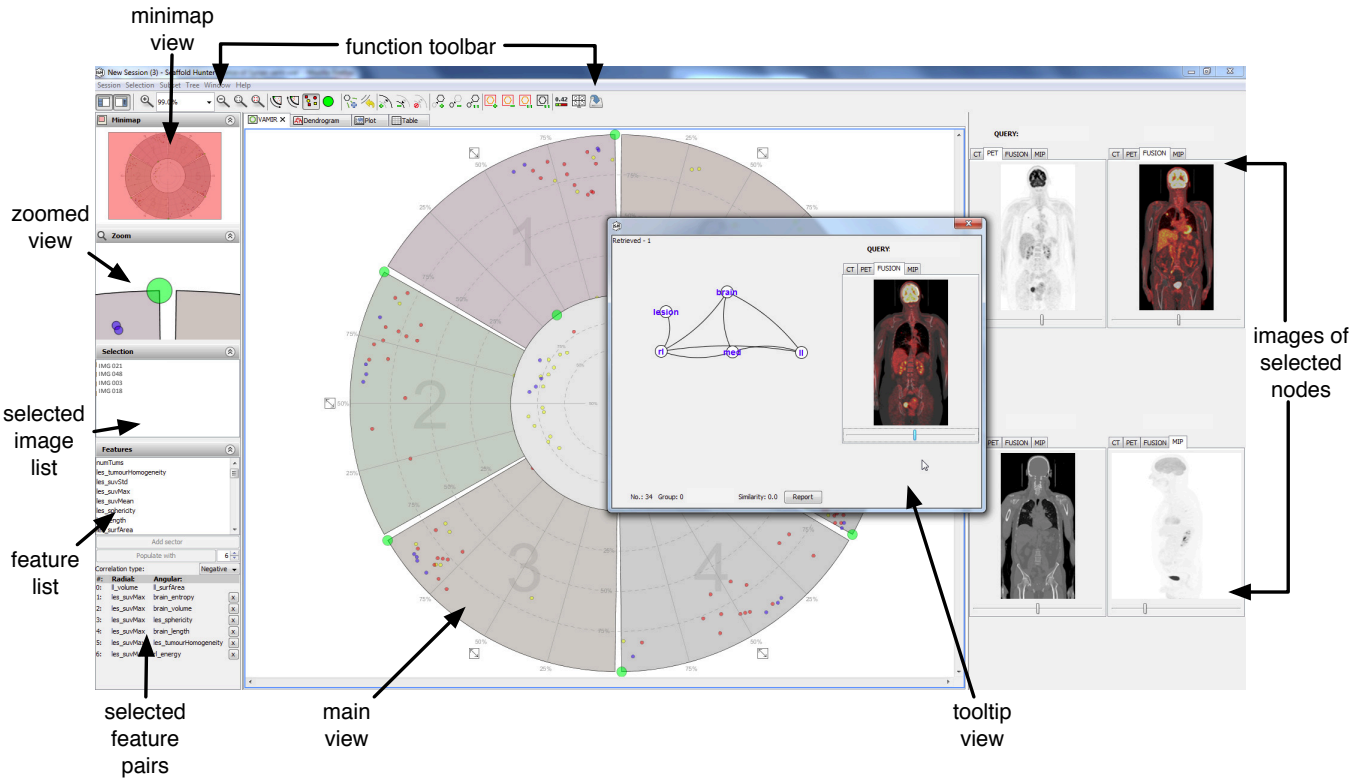

Fig. 2. The VAMIR interface. The major visualisation and interaction components of the tool have been labelled; volumes displayed are coronal and sagittal PET-CT, fused PET-CT, and projection images from a PET-CT scan.

are plotted according to their difference in the feature values. The coordinates of each node are determined as follows.

Let n be the number of sectors (number of feature subsets), r be the radius of the super-centre, and R the radius of the outer divisions[1]. Also let f_1 and f_2 be two image features. Let q_{f_1} and q_{f_2} be the normalised values of these features in the query image. Similarly, let s_{f_1} and s_{f_2} be the normalised values of these features in a database image. Then the polar coordinates (p, θ) for plotting the node are given by:

$$p = r + (1 - |q_{f_1} - s_{f_1}|)(R - r) \tag{1}$$

and

$$\theta = \frac{2\pi}{n}(|q_{f_2} - s_{f_2}| - c + 1) \tag{2}$$

where $1 \leq c \leq n$ is the sector in which the pairing is to be plotted. In the super-centre, the nodes are plotted using the dissimilarity value d calculated by the CBIR engine. The polar coordinates (p_c, θ_c) are given by:

$$p_c = r\left(1 - \frac{d}{\max(D)}\right) \tag{3}$$

and

$$\theta_c = 4\pi\sigma_f \tag{4}$$

where D is the set of dissimilarity values of all images in the database and σ_f is the standard deviation of all the features in an image.

[1] R and r are based upon the size of the display panel with the constraint that $R > r > 0$.

In addition to the position, the shape and color of each node also conveys information to the user. Circles represent the case where the first feature value is larger than or equal to the query's feature first value, i.e., $s_{f_1} \geq q_{f_1}$ (in a sector) or $d \geq 0$ (in the super-centre). Squares represent the alternate case where $s_{f_1} < q_{f_1}$. Color represents similar characteristics for the second feature value: green indicates the query node, red indicates that $s_{f_2} \geq q_{f_2}$ (sector) or $s_{\sigma_f} > q_{\sigma_f}$ (super-centre), and yellow indicates that $s_{f_2} < q_{f_2}$ (sector) or $s_{\sigma_f} < q_{\sigma_f}$ (super-centre). Here, s_{σ_f} is the standard deviation of the normalised features in a particular database image while q_{σ_f} is the standard deviation of the normalised features in the query image. Blue nodes indicate those that have been selected (see Section IV-C for more on selection). The various node characteristics are depicted in Fig. 4.

B. Justification for Circular Visualisation

Our decision to choose a circular layout over a grid-based layout was based on a number of considerations:

1) A circular layout places sectors in equivalent positions while a grid might undesirably imply varying significance of the pairings based on ordering.
2) A circular layout enables smooth dynamic changes, since adding or deleting sectors will not affect the layout's logical consistency.
3) In each sector, the nodes representing the most relevant images are visualised at a high radial (p) value where the angular resolution is better.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

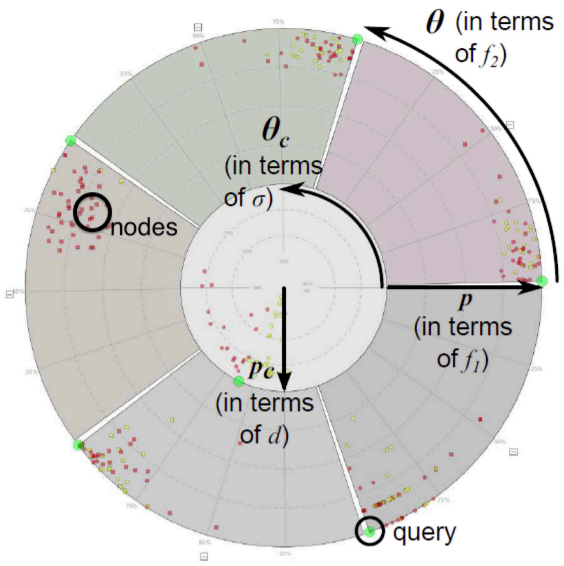

Fig. 3. The main visualisation in VAMIR. Each sector represents the distribution of images compared to the query on two features, f_1 (visualised by the radial coordinate p) and f_2 (visualised by the angular coordinate θ) . The super-centre represents the distribution of the image similarity calculated by the CBIR algorithm, d (visualised by the radial coordinate p_c), and the standard deviation of image features, σ (visualised by the angular coordinate θ_c). Each node represents an image. Groups of nodes in close proximity indicate images with similar pairs of features.

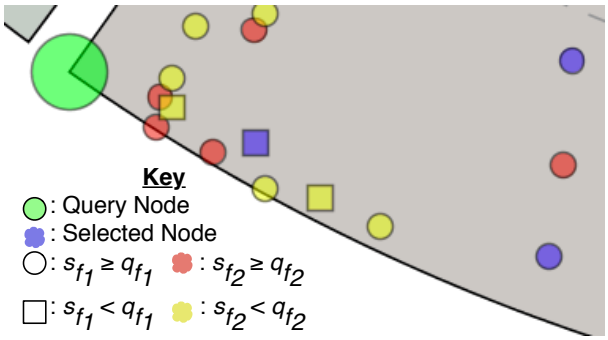

Fig. 4. The different node types in VAMIR. In the key, q is the query image, s is a retrieved image, f_1 is the feature on the radial axis of the sector, and f_2 is the feature on the angular axis.

4) It is possible to position a central disk (the super-centre) amidst the other sectors.

Also the super-centre allows us to visualise the overall similarity computed by the CBIR engine (radial component) and the compactness, i.e., standard deviation of all features of an image (angular component). This allows the user to compare image similarity calculated using all features (the super-centre) to image similarity calculated using a subset of features (the outer sectors).

C. Feature Selection

We provided the ability to change the number of sectors (n) and the feature pairings visualised in each sector. As common CBIR algorithms can employ hundreds of features, it is not feasible to display all possible feature combinations. Only a few features are usually relevant [32] for an explanation of the similarity values and the analysis of outliers. Three methods were used for selecting the feature pairings to be visualised.

1) Automatic Feature Selection: Our default selection of feature pairings was for situations where the user does not have *a priori* information about the data. This feature selection was based on finding pairings that have high absolute correlation values. Correlation values serve as an intuitive indicator for finding feature combinations as they are closely connected to similarity-based classification tasks [32], [33]. Note that in VAMIR, the feature selection is for the visualisation and *not* for the initial retrieval, where a high correlation between two features would hint at a possible redundancy. VAMIR enables a user to analyse the feature space after the querying process, where the visualisation of highly-correlated pairings might be helpful in discovering links between relevant image features. Automatic feature selection is performed by calculating the Pearson correlation coefficient, r_{cor}, for each feature pairing (f_1, f_2), and selecting the pairs with the highest absolute correlation values.

2) Semi-Automatic Selection: We implemented a semi-automatic feature selection method for the case where the user is interested in one specific feature but is unsure of what pairings would be appropriate. In this scenario, the user selects a feature and VAMIR automatically derives the n most positively or negatively correlated features. Each sector then plots the user selected feature f_1 alongside one of the correlated features (f_2).

3) Manual Selection: The user manually specifies feature pairings based on their domain knowledge and prior experience for the particular task. In the user interface, two features can be chosen from a complete list of all available features to populate a new sector at any time. A dynamic table (see Fig. 2) explains the current feature mappings for each sector and provides buttons for removing some or all of the sectors.

IV. Visual Analytics Interactions

We introduced several interactions to enable the user to explore the database, compare the differences in image features, and visually inspect individual images. These interactive capabilities are described in the following subsections.

A. Image Visualisation

While the VAMIR visualisation provides new insights into the similarity of images, it is still necessary for a user to view the actual image to judge the overall characteristics of a case (e.g., disease stage). A link between the abstract visualisation and the actual image is provided to verify if an image discovered through the exploration of the feature space matches the given search intent. This link is established in VAMIR through two approaches.

As shown in Fig. 2 (tooltip view), hovering over a node with the mouse cursor displays a pop-up window containing multiple views of the image associated with the node. The figure depicts the pop-up window when the node represents a PET-CT volume. Using tab controls, the user may alternately view the window contains the PET and CT volumes, a fused volume, and a 3D projection of the PET intensities.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

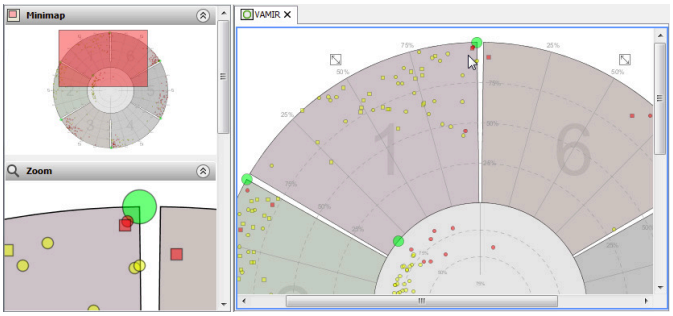

Fig. 5. The minimap provides an overview of the entire visualisation while the "zoom" window shows a magnified view of the area around the mouse cursor. The user can also change the magnification of the main view to examine several levels of detail.

Supplementary information can also be presented within this window. In Fig. 2 this supplementary information takes the form of a graph abstraction that describes the arrangement of tumors and anatomy in the image, which is useful for the interpretation of PET-CT volumes [34]. The pop-up window can be repositioned to avoid obscuring the visualisation.

The second approach, also shown in Fig. 2, is a dynamic grid-based image display in the right sidebar. The query image is always shown in the top-left while the remaining panels react to the user's node selection. Up to three selected images can be visually compared to the query, to each other, and to any pop-up windows. Each panel allows the user to choose from the different representations described above. This image display approach offers the possibility to compare multiple images within VAMIR.

B. Minimap and Zoom

In our visualisation node proximity is an indicator of image similarity and thus it may be necessary for users to examine the visualisation at several levels of detail to separate tightly packed nodes or, alternately, to view the entire spectrum of nodes. Zooming continuously in and out of the main view can be done using the mouse wheel as shown in Fig. 5. A minimap constantly presents the user with a view of the entire canvas, providing a summary of the distributions of feature pairs in different sectors. Similarly, a magnified view of the area around the mouse cursor is always shown on screen to enhance the user's ability to quickly distinguish and visually separate nodes that are clustered closely together.

C. Node Selection

Our visualisation is based on pairings of single features to allow users to examine the contribution of individual features to the image similarity. We use a global selection scheme to allow the user to compare images using multiple groups of features. The global selection links the nodes across all sectors and the super-centre. When a user selects nodes in one sector, the corresponding nodes in all other sectors and in the super-centre are also selected. Selecting multiple nodes in one sector allows the user to discover if images are clustered near each other in all the feature pairings (high overall similarity) or only in some of them (an outlier).

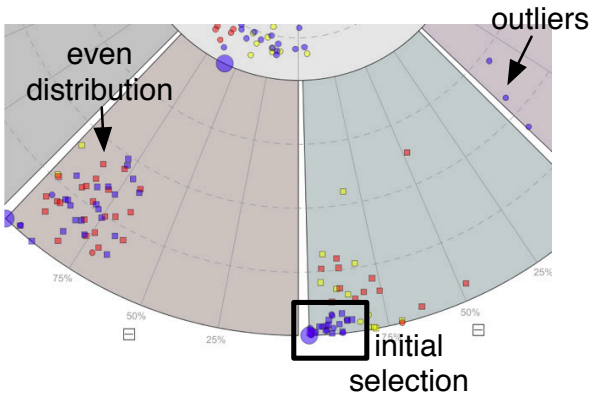

Fig. 6. Selecting nodes allows comparison of images across different feature pairings. Images that are grouped closely together in one pairing may be evenly distributed in another pairing or may be outliers in other pairings.

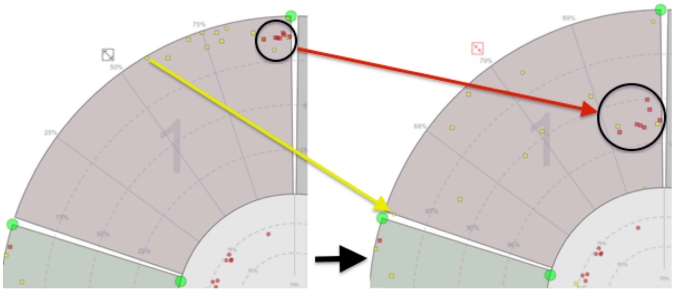

Fig. 7. A rescaled visualisation plots the nodes while using all available space within a sector. The red arrow indicates a cluster of red nodes which are more spread out in the rescaled version. The yellow arrow indicates the position of the node furthest from the query in both the original and rescaled sectors.

The use of selection is shown in Fig. 6. The blue nodes initially selected by the user are closely clustered in one of the sectors. The even distribution of nodes in another sector indicates that the feature pairing of that sector is not important in determining image similarity. In the third sector, several of the selected nodes appear as outliers. The super-centre reveals that some of the selected nodes are further away from the query node, indicating that the CBIR algorithm has considered them to be less similar.

D. Dynamic Querying

When the original query is an outlier compared to the dataset, the user may find it useful to refine the query using a dataset image, whose node is close to the query node in a particular sector or the super-centre. VAMIR then retrieves the similarity of the dataset compared to the new query from the CBIR engine. This similarity information is used to instantly repopulate the visualisation, leaving the feature pairings unchanged. The panels displaying the query and the retrieved images are then updated.

E. Rescaling

When feature values are narrowly distributed, all nodes will be placed close to the query in the corresponding sector, resulting in large sections of empty space. We allow the user

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

to fully utilise this empty space through sector scaling in the angular and radial directions. When scaling is performed, the node distribution is linearly stretched across the whole sector as shown in Fig. 7. This helps to identify individual nodes that were initially clustered together and emphasises distances that were difficult to separate previously, e.g., the red nodes in Fig. 7. Rescaling can be triggered sector-wise or across all sectors.

We use several indicators to ensure that a user is aware that a sector has been rescaled. First, the rescaling process is applied using a short animation, with nodes moving to their new positions. In addition, the sector's axis annotations indicating the percentage of similarity are updated to reflect the new range. The color and arrow symbol on the rescaling button for the sector is also changed.

V. Datasets

A. PET-CT Lung Cancer Dataset

We used 50 PET-CT volumes of lung cancer patients acquired on a Siemens Biograph mCT scanner with a PET resolution of 200×200 pixels at 4.07mm^2, a CT resolution of 512×512 pixels at 0.98mm^2, and a slice thickness of 3mm. Each PET and CT volume contained 326 slices on average, for a total of 32600 2D images across the dataset. The images contained between 1 to 7 tumors with a variety of different spatial and localisation properties. The clinical reports, which contained details about the primary lung tumor and the presence or absence of nodal involvement were also available. We used the tumor and nodal properties as labels for the images; these labels were the ground-truth for validating image similarity. Images that were given at least one label in common were considered to be similar. All data used in the study were de-identified.

We used a well-established adaptive thresholding algorithm with refinements to segment the lung ROIs from the CT [35]. Tumors from the PET volumes were segmented with a 40% connected thresholding at the positions with locally peak standard uptake values (SUVs), the normalised value of the PET radiotracer uptake [36]. We applied manual connected thresholding to coarsely segment the brain and mediastinal tissue (including the heart) to include major organs above the diaphragm. We used the findings from the clinical reports to make minor corrections to the segmented ROIs to ensure that the segments were well-defined prior to automatic feature extraction as described in Section VI-B.

B. LIDC/IDRI

We also used a large public CT dataset. The LIDC/IDRI dataset contains 1018 CT volumes, with over 240,000 2D slices [37].[2] Each CT volume had a pixel resolution of 512×512 pixels with a varying numbers of slices (average 242, minimum 65, maximum 1056 slices). The slice thickness ranged from 0.6mm to 5mm. The lung nodules within these volumes were delineated by several experts and were available as a reference.

[2]The LIDC/IDRI dataset can be obtained from http://cancerimagingarchive.net/.

VI. Implementation

A. VAMIR Visualisation and Interactions

VAMIR was implemented as an extension of Scaffold Hunter [38], a tool for the interactive visual analysis of chemical compound databases. We selected Scaffold Hunter for its extensive VA library, which allows for multiple linked views and provides techniques for clustering and plotting data within a fluent (smooth) visualisation. Scaffold Hunter also has a modular design that includes a flexible plugin architecture. Many of its core features for visualisation, view organisation, data management and analysis are similar to VAMIR's requirements, making it an appropriate choice as a framework for our implementation. Scaffold Hunter is written in Java, so VAMIR would be available on most common platforms.

We extended the functionality of Scaffold Hunter by implementing the new visualisation described in Section III and the interaction capabilities in Section IV. Hence, VA was applied to medical images instead of the scaffolds that represent chemical compounds.

B. PET-CT CBIR

We used a graph-based CBIR method optimised for the retrieval of PET-CT volumes [15]. Image features were automatically derived from the PET-CT volumes and represented in a relational graph structure. The features used were region properties (size, surface area, etc.), image texture, intensity information, and tumor SUV.

The quantitative comparison of two images was performed by computing the edit distances of their respective graph representations. Refer to Kumar et al. [15] for details about the features and the algorithms that were used. The most similar images are generally expected within the first 5 to 20 retrieved images [12]; we chose 20 to allow further lower ranked images to be considered. This threshold can be adjusted by the user.

C. LIDC/IDRI CBIR

We implemented a texture-based retrieval method for the LIDC/IDRI dataset. This retrieval method used the well-established Haralick texture features [39]. Thirteen gray-level co-occurence matrices (one for each unique 3D direction) of neighbouring voxels were used to calculate five Haralick features: contrast, correlation, energy, entropy, and homogeneity. The features were calculated locally from the nodules. The similarity of two images was computed using the Euclidean distance, a standard method for computing the differences between feature vectors [5].

VII. Evaluation Procedure

We evaluated VAMIR in three ways. Two case studies were used to demonstrate the capabilities of VAMIR. We also conducted a user study to gain qualitative insights about VAMIR's capabilities. Finally, we examined the resource usage and computational performance of VAMIR when using different datasets.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

TABLE I
USER SURVEY

	Statement
1	The viewer was easy to use.
2	The viewer was fast and responsive to input.
3	The viewer rendered good quality images.
4	The viewer provided all the controls I needed.
5	The viewer and its controls were well laid out.
6	I was able to use the viewer to understand the images.
7	The viewer contained all information necessary to understand the images.
8	The amount of image and other information presented did not confuse me.
9	When searching, the viewer generally found the most similar images first.
10	I was able to use the search and filtering framework to improve my understanding of the image data.

A. Case Studies

We conducted two case studies, one for each dataset and CBIR engine, to demonstrate the capabilities of VAMIR. The scenario for our evaluation is the retrieval of images with certain properties for compilation into datasets for use in lung cancer research.

The PET-CT case study (Section VIII-A) examined the retrieval of images where tumor features and relationships were important; the user would have to examine tumor features as well as the location of the tumor in an volume. The LIDC/IDRI case study (Section VIII-B) investigated VAMIR's usefulness when relevant volumes were given low rankings by the ranked list approach; the user would have to isolate the features most relevant to their task and choose the most similar images on the basis of those features. For each case study, we describe the process followed to explore the dataset and the interactions used to select the most similar images. We also show the visualisations produced by VAMIR during the exploration process and explain how these visualisations indicate image similarity or dissimilarity.

B. User Evaluation

Approval was obtained from our institutional human ethics committee for conducting a user evaluation of VAMIR. Since the primary use case for VAMIR identified in this paper was scientific research and education (see Section I), we recruited from our institution 5 researchers engaged in different areas of biomedical research. All the researchers recruited as participants for the user study were otherwise uninvolved with VAMIR.

Each participant was first given a training session with a walkthrough of functionalities provided by VAMIR. The participants were then asked to complete two retrieval tasks, one with each dataset. The scenario for each task was collating similar images into a dataset for use in a scientific study, e.g., finding a subset for retrospective analysis of particular image characteristics, classification, etc. In each task, the participants queried the dataset and used VAMIR's capabilities to decide whether the retrieved images were similar to the query and if there were any feature patterns shared by similar images in the dataset. The decision of which images were similar were left to the subjective choice of the participants since in practice humans would have different interpretations of image similarity based on their own experiences and needs.

At the end of each task, the participants completed an anonymous survey where they were asked about their experiences when using VAMIR. In the survey, the participants indicated their level of agreement or disagreement with several statements using a 5-point Likert scale (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree); a 5-point scale is a well-established approach for gathering information about user experiences [40]. Participants were also allowed to leave free text comments on any aspect of VAMIR.

The statements in the survey are given in Table I. We adapted the standard System Usability Scale (SUS) [41] by including statements that were specific for image search and visualisation applications. Several of the Statements (e.g., 1, 5, and 8) rephrased the SUS statements for our application while others (e.g., 9 and 10) were specifically included to evaluate the users satisfaction about the effectiveness of the system for image search. This survey has also been used in a previous study [34].

C. Performance Evaluation

The evaluation was conducted on a consumer PC with a 2.67 GHz Intel i5 CPU and 4 GB RAM; the operating system was the 64-bit edition of Windows 7 Professional. The CPU utilisation and RAM usage during the user studies (described in Section VII-B above) were monitored and logged. We evaluated the performance of VAMIR by analysing these logs.

VIII. RESULTS

A. Case Study – PET-CT Dataset

The query was an image with a large tumor in the right lung. The focus of the query was on tumor size and so semi-automatic feature selection (see Section III-C2) was used to populate six sectors with tumor volume on the radial axis and the automatically selected features to the angular axes. The initial visualisation is shown in Fig. 8; all 20 of the retrieved results are placed quite far from the query in the radial direction, indicating much smaller tumor sizes, i.e., the query itself was an outlier. The graph representation of the query was compared to the graph of the radially nearest node (most similar tumor size); both cases contained a single tumor shared between the mediastinum and the right lung. Due to the structural and feature similarity, we refined the retrieval process by using this retrieved image as a new query to eliminate the outlier status of the initial query.

The resulting visualisation, shown in Fig. 9, exhibits a more diverse node distribution. Highlighted in Fig. 10 are the seven cases that were most similar to the new query with respect to tumor size. The tumor size feature of these seven images were within 25% of the value of the query (i.e., at worst 75% similarity); five of these seven images also had tumor sphericity that was within 50% similarity to the query.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

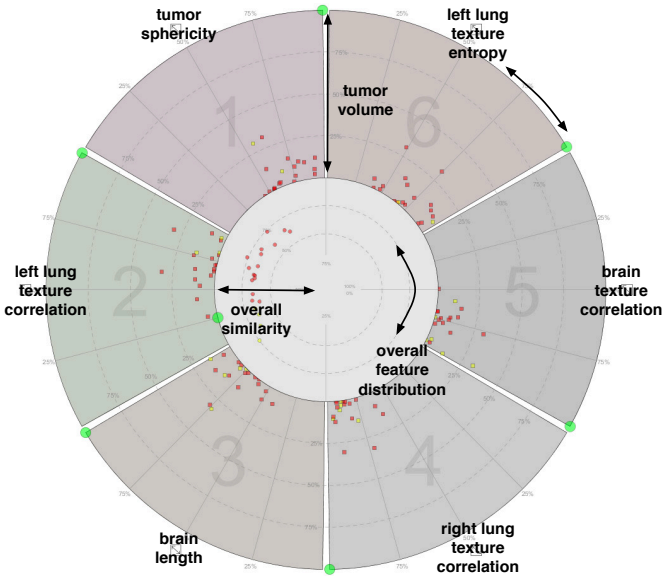

Fig. 8. PET-CT Case Study – The feature space using the original query image.

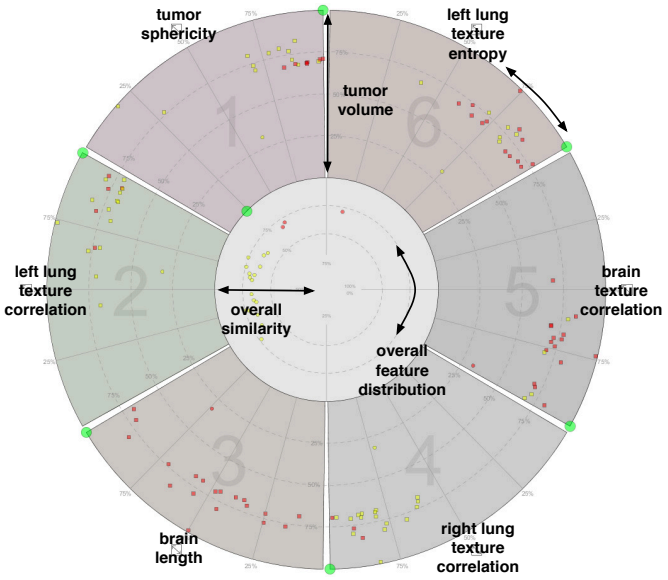

Fig. 9. PET-CT Case Study – Refining the feature space by using one of the retrieved images as a new, refined query.

Analysis of their graph representations revealed that five of these cases had a tumor configuration that was similar to the query (between mediastinum and right lung), while the remaining two contained tumors limited to the right lung.

We then customised the feature selection and paired the lesion size with features describing the right lung and the mediastinum (volume, surface area and tissue homogeneity) to place emphasis on the tumor configuration of the original query. The seven images selected before were all contained in the top twenty results of both queries we conducted.

After visually inspecting all sectors, we selected eight images from the new query as being most similar. The nodes of these eight images were in close proximity to the query node in multiple sectors; the features used included tumor volume

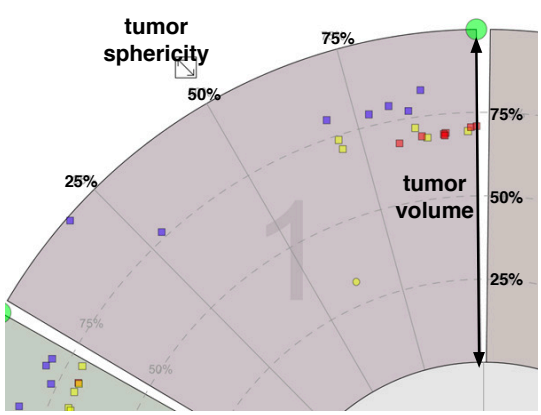

Fig. 10. PET-CT Case Study – The selected (blue) nodes represent the images most similar to the refined query in terms of tumor size.

and sphericity, right lung and mediastinum volume, area, and homogeneity. According to the ground-truth (see Section V-A), six of the selected images (75%) were relevant to the original query. In comparison, only 35% of the images retrieved using a ranked list approach were relevant to the original query. This result indicates VAMIR allowed us to explore the retrieved dataset for the most relevant images.

B. Case Study – LIDC/IDRI Dataset

The query was a CT volume containing six nodules. The query was typical of the images in the dataset and thus most of the images in the dataset had a low distance (high feature similarity) during the retrieval using the CBIR engine. Fig. 11 shows how the distribution of the retrieved images in the super-centre. The proximity of the nodes to the outer edge of the super-centre (i.e., their distance from the query using all features) indicates that a ranked list approach would consider most of these images to be similar to the query.

To locate the images with similar nodule properties, we chose to analyse the retrieved images in terms of the texture homogeneity. VAMIR's semi-automatic feature selection was used to find the features most relevant to homogeneity. After visual inspection, it was immediately apparent that the pairing of homogeneity and contrast features revealed a set of images that were more similar to the query than other images in the dataset. This group of images is represented by the group of nodes covered by the dashed line in Fig. 12.

We visually compared the retrieval using VAMIR and using the ranked list by selecting the similar nodes indicated in Fig. 12. The results of our selection are shown in Fig. 13. The most similar images based on the homogeneity and contrast features are not the most similar images when using all features (i.e., the ranked list). This can be inferred by the radial position (distance from the outer edge) of the selected (blue) nodes in the super-centre. Many selected nodes fall into the middle of the ranked list, indicated by the many unselected nodes that are closer to the query (green) node. This result indicates that VAMIR allowed us to find images that would be given low rankings by conventional approaches.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

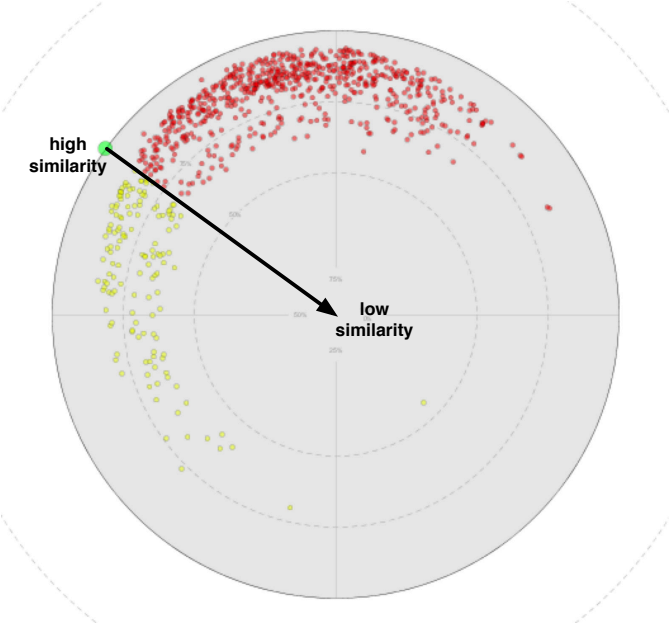

Fig. 11. LIDC/IDRI Case Study – The grouping of retrieved image nodes near the query node in the super-centre implies that the ranked list considers most of the images to be similar to the query.

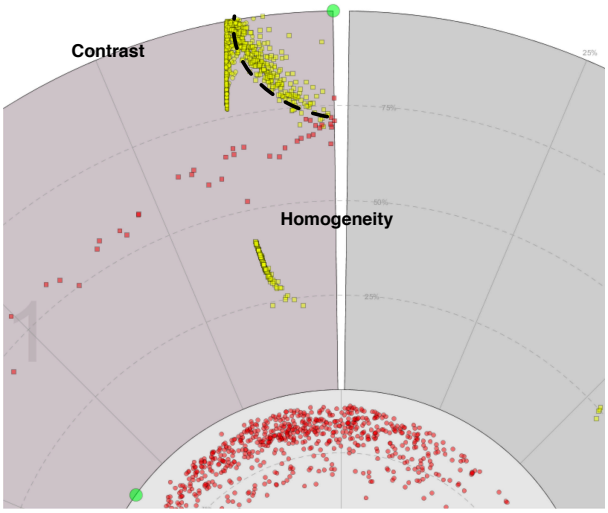

Fig. 12. LIDC/IDRI Case Study – Examining images based on texture homogeneity and contrast reveals a group of images similar to the query (indicated by the dashed line).

C. User Study Survey Responses

Figure 14 summarises the survey responses across all participants. Table II shows the median and range of the responses in our user study. We calculated the p value using the Student's t-test to compare the user evaluation using the PET-CT and LIDC/IDRI datasets. The results show that there were no significant differences ($p > 0.05$) in the responses for all the statements in the survey.

D. Participant Comments

The majority of free text comments given by the participants were suggestions on new capabilities that could be added to tune VAMIR to their individual preferences. The following is

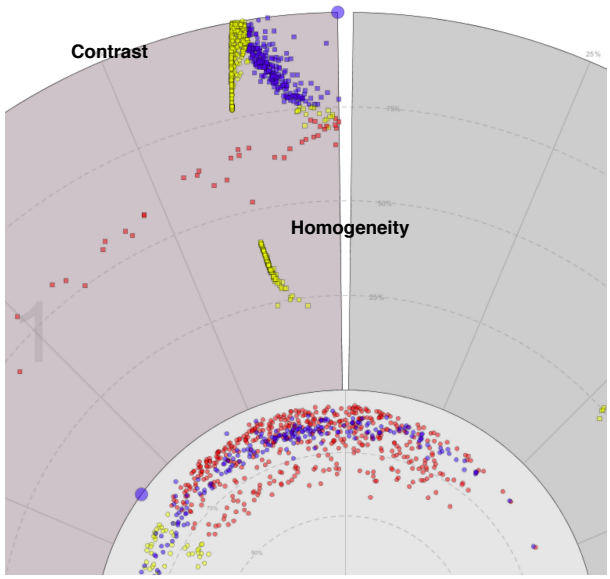

Fig. 13. LIDC/IDRI Case Study – The selected (blue) nodes represent the most similar images but have lower rankings in the ranked list as seen by their position in the super-centre.

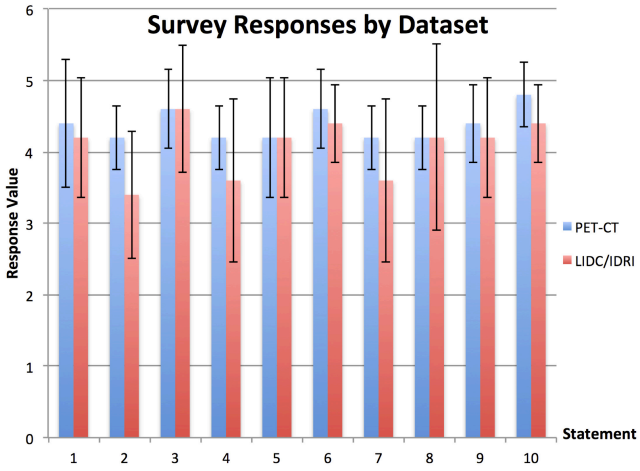

Fig. 14. The mean responses to survey questions across all participants; error bars indicate the standard deviation of the responses.

TABLE II
SURVEY RESPONSE SUMMARY

Statement	Scores				p value
	LIDC		PET-CT		
	Median	Range	Median	Range	
1	4	3—5	5	3—5	0.7245
2	4	2—4	4	4—5	0.1248
3	5	3—5	5	4—5	1.0000
4	4	2—5	4	4—5	0.3214
5	4	3—5	4	3—5	1.0000
6	4	4—5	5	4—5	0.5796
7	4	2—5	4	4—5	0.3214
8	5	2—5	4	4—5	1.0000
9	4	3—5	4	4—5	0.6684
10	4	4—5	5	4—5	0.2429

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

TABLE III
PERFORMANCE EVALUATION SUMMARY

	RAM (MB)		CPU (%)	
	Mean	Peak	Mean	Peak
PET-CT	247	473	3.5	27.0
LIDC	205	535	4.8	22.5
p value	0.2325	0.5451	0.1365	0.7033

a summary of the suggestions made by the participants and the number of participants who made similar comments:

- The tool should integrate text and numerical information for each image, either as a tooltip or in the side panel (3 comments).
- There should be more node visualisation options, e.g., varying shapes, colors, opacity (2 comments).
- There should be a shortcut for switching between views and features (2 comments).
- The tool should allow users to customise controls and shortcuts for common operations (1 comment).
- There should be a paintbrush selection tool (1 comment).
- The tool should integrate image manipulation tools, e.g. for varying the brightness and contrast (1 comment).
- There should be a visual link that depicts the correspondence between selected nodes and the images displayed in the side-bar (1 comment).

The participants noted the following limitations:

- Image load time and slice-based views could be faster (2 comments).
- At times it could be confusing to realise which position in a sector was most related. (1 comment).

We also received the following positive comment about VAMIR: the tool was good for interpreting the relationships in a large dataset and determining the distribution of patients or images (2 comments).

E. Computational Performance and Scalability

Table III summarises the CPU and RAM usage of VAMIR, across all runs during the user study. We compared the resource usage when using the smaller PET-CT dataset and the larger LIDC/IDRI dataset; the table also shows the p value of the difference calculated using the Student's t-test. The results show that VAMIR had an average CPU utilisation of 4.8%, and did not exceed 27% of the CPU and 535 megabytes of memory usage. There was no significant difference in the resource usage between the two datasets ($p > 0.05$). This result is consistent with the survey responses regarding performance (see Statement 2 in Fig. 14 and Table II).

IX. DISCUSSION

Our findings show that VAMIR can be used to retrieve images by enabling a guided visual exploration of the entire dataset. In each of the case studies, VAMIR allowed the images in the dataset to be interpreted based on pairs of features using the query as a point-of-reference. This enabled users to compare the similarity of specific features in the query image and the images in the dataset. Users could then select the relevant images based on the intent of their search.

VAMIR also allowed users to locate images that were missed or given low ranks by traditional approaches. This is best depicted in the case study using the LIDC/IDRI dataset (see Section VIII-B and Fig. 13). A group of relevant images, where the texture contrast and texture homogeneity were similar to that of the query, was given a low ranking in the super-centre (which mimics the ranked list). Finding all the lowly ranked images using traditional ranked list approaches would be impractical since users would need to first view and discard dozens of irrelevant CT volumes with high rankings. The impracticality of the ranked list approach is further increased by the complexity of the images; high resolution CT volumes with hundreds of slices could potentially take a much longer time to inspect and interpret. VAMIR reduces the need to inspect the irrelevant images by redirecting the users to the most relevant images according to the features they chose.

In our user study, VAMIR achieved a high average rating (≥ 4 out of 5) for a majority of the survey statements. This indicates that VAMIR was a useful tool for examining the entire dataset and allowing users to subjectively select the most similar images to a query. VAMIR acts to reduce the semantic gap by allowing users to subjectively specify the images they consider to be similar, which depends upon the users' prior experiences, the data they are examining, and the task they are performing. The high responses (≥ 4) to Statements 9 and 10 of the survey indicate that the participants were able to use VAMIR to find images that they considered to be relevant.

Statements 2, 4, and 7 scored the lowest responses in the survey. On average the participants were neutral about these statements (score ≥ 3) for the LIDC/IDRI dataset. The median and range of the responses (in Table II) indicates that some of the participants agreed (score = 4) or strongly agreed (score = 5) with these statements. The differences in these statements can possibly be explained by how the individual participants used the system.

The lower responses for Statement 2 regarding the speed of the system are most likely due to participants viewing the volumes associated with the nodes in the VAMIR visualisation (see Section IV-A). Time is needed to load individual CT volumes from the LIDC/IDRI dataset or the PET-CT volumes from the PET-CT dataset. The scores for Statement 2 were lower for the LIDC/IDRI dataset because the volumes were larger (≈ 120 MB on average) with a corresponding increase in load times. We suggest that the participants who scored Statement 2 highly (≥ 4) relied mainly upon the VAMIR feature visualisation and inspected fewer volumes directly while those who gave it a lower score (≤ 3) would have directly viewed more volumes. Our reason for the low score is validated by the participant comments (see Section VIII-D) in which the slow volume load time was noted as a limitation.

Statements 4 and 7 are related to the information presented and the controls provided by VAMIR. The median and range of the responses indicates that most users agreed with these statements. The lower scores by some participants are expected since the individual participants would have quite different needs. The participant comments (Section VIII-D) indicate the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

areas where participants would customise the existing capabilities of VAMIR for their particular needs, such as adapting VAMIR's visualisation capabilities (e.g., having more node shapes, colors, etc.), integrating alternate forms of information (e.g., numerical and text), or allowing the user to customise the interactions (e.g., controls, view switching, etc.).

Our user study also compared the differences in VAMIR when using a small PET-CT dataset and a larger LIDC/IDRI dataset. As shown in Table II, there were no significant differences ($p > 0.05$) in the responses we received when evaluating VAMIR with the different datasets. Based on the high responses (≥ 4) to Statements 9 and 10 we can conclude that the choice of dataset did not have a significant impact in the usefulness of VAMIR, i.e., the users thought that VAMIR was useful for both the datasets. The performance results (Section VIII-E) show that there were no significant differences ($p > 0.05$) in the CPU utilisation and RAM consumption when using the different datasets. The explanation for this is that VAMIR processes the feature vectors and similarity scores computed by the CBIR engine without any need to process the image data directly. The image visualisation capabilities of VAMIR (as described in Section IV-A) loads the images into memory only if requested by a user interaction. However, this reduction in overall resource consumption comes at the cost of increased image load times when the user requests multiple images at the same time. This trade-off is necessary as it enables VAMIR to scale to large datasets of volumetric images. Traditionally, most CBIR systems only present between 5 to 20 most relevant images [12]. VAMIR's scalability means that it is possible to consider hundreds of data points at once. Users are able to select the most relevant images from the larger dataset based on their own understanding of the feature space and how the retrieved images relate to the query.

VAMIR allows users to retrieve images by exploring the entire dataset based on the features that they consider to be important. By considering the entire dataset, users are able to discover images that may be missed or given low rankings by methods that use a ranked list. VAMIR differs from other approaches for identifying lowly ranked images in several ways. Unlike techniques based on automatic feature selection, VAMIR does not require training which means that it does not have to be retrained for new datasets or new clinical application domains. Users may instead choose features based upon their prior knowledge of the application domain or may elect to use a correlation-based feature selection algorithm that suggests the features that are most likely to be relevant. An interactive feature selection means that query-based feature selection does not rely upon hierarchical classification, as in the approach reported by Dy et al. [18], where an incorrect classification at one level of the hierarchy will cascade to suboptimal feature selections at other levels.

The feature selection in VAMIR is for the purpose of visualisation. The choice of features in VAMIR does not impact the results of the CBIR engine (displayed in the super-centre) but instead provides complementary views of the dataset in terms of the selected features. Since the choice of features does not impact the results from the CBIR engine, the user does not need to understand the underlying retrieval algorithm to reach a satisfactory outcome. VAMIR's approach reduces the burden of technical knowledge from the user. This is in contrast to weighting-based techniques for discovering lowly ranked images in which a user must understand the relationship between feature weights and the retrieval algorithm [21].

In contrast to relevance feedback, VAMIR allows the user to select the most relevant images from the entire dataset instead of a subset of the ranked list. Furthermore, while relevance feedback trains classifiers for feature selection [13], [20], VAMIR gives the user direct control over the features to be examined. This allows the user to define a specific feature set based on their domain knowledge. Relevance feedback is well suited to 2D imaging where it is relatively efficient to visualise and compare dozens of images simultaneously. We suggest that VAMIR's approach is more optimised for volumetric and multi-modality imaging, where individual scans consist of hundreds or thousands of 2D images.

X. Conclusions

In this study, we introduced Visual Analytics for Medical Image Retrieval (VAMIR) as a new CBIR framework. Our evaluation showed that VAMIR's novel approach to visual exploration facilitated interactive human exploration of the retrieved results derived from an automated image similarity matching. The capabilities of VAMIR allow users who may not be experts in retrieval algorithms to explore the feature space with assistance from the retrieval process to discover the images that they consider similar, thereby reducing the semantic gap.

Our future work will involve using relevance feedback to complement the VA techniques in VAMIR. We also plan to: a) conduct usability studies and investigations of how VAMIR can be customised for different clinical user groups, through the introduction of domain-specific visualisations and interactions; b) investigate methods for maintaining the usefulness of VAMIR as an interactive tool as clinical datasets expand, for example by visualising very similar nodes as groups or hierarchies; c) examine how interactivity can be maintained for large feature sets for which the computation of pairwise feature correlations is time consuming; and d) use VAMIR to discover the relationships between visual features, semantic annotations, and clinical attributes.

Acknowledgments

The authors are grateful for the direct and indirect contributions of our collaborators at the Royal Prince Alfred Hospital, Sydney, Australia.

The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study.

The authors are grateful for the support given to F. Nette by Google Summer of Code.

The authors wish to thank the anonymous reviewers, whose comments helped us to significantly improve the presentation of our contribution.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2014.2361318, IEEE Journal of Biomedical and Health Informatics

REFERENCES

[1] H. Müller, J. Kalpathy-Cramer, B. Caputo, T. Syeda-Mahmood, and F. Wang, "Overview of the first workshop on medical content–based retrieval for clinical decision support at MICCAI 2009," in *Lect Notes Comput Sc*, 2010, vol. 5853, pp. 1–17.

[2] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," *Int J Med Inform*, vol. 73, no. 1, pp. 1 – 23, 2004.

[3] S. A. Napel, C. F. Beaulieu, C. Rodriguez, J. Cui, J. Xu, A. Gupta, D. Korenblum, H. Greenspan, Y. Ma, and D. L. Rubin, "Automated retrieval of CT images of liver lesions on the basis of image similarity: Method and preliminary results," *Radiology*, vol. 256, no. 1, pp. 243–252, 2010.

[4] H. Müller, A. Rosset, A. Garcia, J.-P. Vallée, and A. Geissbuhler, "Benefits of content-based visual data access in radiology," *Radiographics*, vol. 25, no. 3, pp. 849–858, 2005.

[5] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE T Pattern Anal*, vol. 22, no. 12, pp. 1349 –1380, 2000.

[6] T. Deserno, S. Antani, and R. Long, "Ontology of gaps in content-based image retrieval," *J Digit Imaging*, vol. 22, no. 2, pp. 202–215, 2009.

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput Surv*, vol. 40, no. 2, pp. 5:1–5:60, 2008.

[8] L. R. Long, S. Antani, T. M. Deserno, and G. R. Thoma, "Content-based image retrieval in medicine: Retrospective assessment, state of the art, and future directions." *International journal of healthcare information systems and informatics*, vol. 4, no. 1, pp. 1–16, 2009.

[9] C.-R. Shyu, C. E. Brodley, A. C. Kak, A. Kosaka, A. M. Aisen, and L. S. Broderick, "ASSERT: A physician-in-the-loop content-based retrieval system for HRCT image databases," *Comput Vis Image Und*, vol. 75, no. 1-2, pp. 111–132, 1999.

[10] G. Quellec, M. Lamard, G. Cazuguel, B. Cochener, and C. Roux, "Wavelet optimization for content-based image retrieval in medical databases," *Med Image Anal*, vol. 14, no. 2, pp. 227–241, 2010.

[11] G. Quellec, M. Lamard, L. Bekri, G. Cazuguel, C. Roux, and B. Cochener, "Medical case retrieval from a committee of decision trees," *IEEE T Inf Technol B*, vol. 14, no. 5, pp. 1227–1235, 2010.

[12] G. Quellec, M. Lamard, G. Cazuguel, C. Roux, and B. Cochener, "Case retrieval in medical databases by fusing heterogeneous information," *IEEE T Med Imaging*, vol. 30, no. 1, pp. 108–118, 2011.

[13] T. Deserno, M. Güld, B. Plodowski, K. Spitzer, B. Wein, H. Schubert, H. Ney, and T. Seidl, "Extended query refinement for medical image retrieval," *J Digit Imaging*, vol. 21, no. 3, pp. 280–289, 2008.

[14] W. Hsu, S. Antani, L. R. Long, L. Neve, and G. R. Thoma, "SPIRS: a web-based image retrieval system for large biomedical databases," *Int J Med Inform*, vol. 78, no. Supplement 1, pp. S13–S24, 2009.

[15] A. Kumar, J. Kim, L. Wen, M. Fulham, and D. Feng, "A graph-based approach for the retrieval of multi-modality medical images," *Med Image Anal*, vol. 18, no. 2, pp. 330–342, 2014.

[16] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[17] S. F. da Silva, M. X. Ribeiro, J. d. E. S. Batista Neto, C. Traina-Jr, and A. J. M. Traina, "Improving the ranking quality of medical image retrieval using a genetic feature selection method," *Decis Support Syst*, vol. 51, no. 4, pp. 810–820, 2011.

[18] J. G. Dy, C. E. Brodley, A. Kak, L. S. Broderick, and A. M. Aisen, "Unsupervised feature selection applied to content-based retrieval of lung images," *IEEE T Pattern Anal*, vol. 25, no. 3, pp. 373 – 378, 2003.

[19] W. Jiang, G. Er, Q. Dai, and J. Gu, "Similarity-based online feature selection in content-based image retrieval," *IEEE T Image Process*, vol. 15, no. 3, pp. 702–712, 2006.

[20] M. M. Rahman, P. Bhattacharya, and B. C. Desai, "A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback," *IEEE T Inf Technol B*, vol. 11, no. 1, pp. 58–69, 2007.

[21] X. Xu, D.-J. Lee, S. Antani, and L. Long, "A spine x-ray image retrieval system using partial shape matching," *IEEE T Inf Technol B*, vol. 12, no. 1, pp. 100 –108, . 2008.

[22] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 2, no. 1, pp. 1–19, 2006.

[23] J. Thomas and K. Cook, "A visual analytics agenda," *IEEE Comput Graph*, vol. 26, no. 1, pp. 10–13, 2006.

[24] A. Hiroike, Y. Musha, A. Sugimoto, and Y. Mori, "Visualization of information spaces to retrieve and browse image data," in *Lect Notes Comput Sc*, 1999, vol. 1614, pp. 155–163.

[25] Y. Gao, C. Yang, Y. Shen, and J. Fan, "Incorporate visual analytics to design a human-centered computing framework for personalized classifier training and image retrieval." in *Advances in Information and Intelligent Systems*, ser. Studies in Computational Intelligence. Springer, 2009, vol. 251, pp. 165–187.

[26] J. Rodrigues, L. Romani, A. Traina, and C. Traina, "Combining visual analytics and content based data retrieval technology for efficient data analysis," in *IEEE Int Conf Inf Vis*, London, United Kingdom, 2010, pp. 61–67.

[27] O. de Rooij and M. Worring, "Active bucket categorization for high recall video retrieval," *IEEE T Multimedia*, vol. 15, no. 4, pp. 898–907, 2013.

[28] D. Heesch, "A survey of browsing models for content based image retrieval," *Multimed Tools Appl*, vol. 40, no. 2, pp. 261–284, 2008.

[29] F. Nette, A. Kumar, K. Klein, J. Kim, and H. Handels, "VAMIR - Visual Analytics for Medical Image Retrieval: Preliminary Study on PET-CT data," in *Student Conference Medical Engineering Science: Proceedings*, Lübeck, Germany, 2013, pp. 127–130.

[30] M. Tory and T. Moller, "Human factors in visualization research," *IEEE T Vis Comput Gr*, vol. 10, no. 1, pp. 72 –84, 2004.

[31] M. L. Wilson, "Search user interface design," *Synthesis Lectures on Information Concepts, Retrieval, and Services*, vol. 3, no. 3, pp. 1–143, 2011.

[32] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[33] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval: an experimental comparison," *Information Retrieval*, vol. 11, no. 2, pp. 77–107, 2008.

[34] A. Kumar, J. Kim, L. Bi, M. Fulham, and D. Feng, "Designing user interfaces to enhance human interpretation of medical content-based image retrieval: application to PET-CT images," *Int J Comput Assist Rad Surg*, vol. 8, no. 6, pp. 1003–1014, 2013.

[35] S. Hu, E. Hoffman, and J. Reinhardt, "Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images," *IEEE T Med Imaging*, vol. 20, no. 6, pp. 490–498, 2001.

[36] J. Bradley, W. L. Thorstad, S. Mutic, T. R. Miller, F. Dehdashti, B. A. Siegel, W. Bosch, and R. J. Bertrand, "Impact of FDG-PET on radiation therapy volume delineation in non-small-cell lung cancer." *Int J Radiat Oncol Biol Phys*, vol. 59, no. 1, pp. 78–86, 2004.

[37] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P.-Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. Vande Casteele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, 2011.

[38] S. Wetzel, K. Klein, S. Renner, D. Rauh, T. I. Oprea, P. Mutzel, and H. Waldmann, "Interactive exploration of chemical space with Scaffold Hunter," *Nat Chem Biol*, vol. 5, no. 8, pp. 581–583, 2009.

[39] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE T Syst Man Cyb*, vol. 3, no. 6, pp. 610–621, 1973.

[40] E. L.-C. Law, V. Roto, M. Hassenzahl, A. P. Vermeeren, and J. Kort, "Understanding, scoping and defining user experience: a survey approach," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Boston, USA, 2009, pp. 719–728.

[41] J. Brooke, "SUS: a 'quick and dirty' usability scale," in *Usability Evaluation in Industry*, P. W. Jordan, B. A. W. Thomas, and I. L. McClelland, Eds. Taylor & Francis, 1996, pp. 189–194.