Eurographics Workshop on Visual Computing for Biology and Medicine (2016)
S. Bruckner, B. Preim, and A. Vilanova (Editors)

Visual Analytics for the Exploration and Assessment of Segmentation Errors

R.G. Raidou[1,5], F.J.J. Marcelis[1], M. Breeuwer[1,2], M.E. Gröller[3,4], A. Vilanova[5,1], H.M.M. van de Wetering[1]

[1]Eindhoven University of Technology, The Netherlands , [2]Philips Healthcare Best, The Netherlands
[3]Institute of Computer Graphics and Algorithms, Vienna University of Technology, Austria
[4]VRVis Research Center, Austria, [5]Delft University of Technology, The Netherlands

Abstract

Several diagnostic and treatment procedures require the segmentation of anatomical structures from medical images. However, the automatic model-based methods that are often employed, may produce inaccurate segmentations. These, if used as input for diagnosis or treatment, can have detrimental effects for the patients. Currently, an analysis to predict which anatomic regions are more prone to inaccuracies, and to determine how to improve segmentation algorithms, cannot be performed. We propose a visual tool to enable experts, working on model-based segmentation algorithms, to explore and analyze the outcomes and errors of their methods. Our approach supports the exploration of errors in a cohort of pelvic organ segmentations, where the performance of an algorithm can be assessed. Also, it enables the detailed exploration and assessment of segmentation errors, in individual subjects. To the best of our knowledge, there is no other tool with comparable functionality. A usage scenario is employed to explore and illustrate the capabilities of our visual tool. To further assess the value of the proposed tool, we performed an evaluation with five segmentation experts. The evaluation participants confirmed the potential of the tool in providing new insight into their data and employed algorithms. They also gave feedback for future improvements.

Categories and Subject Descriptors (according to ACM CCS): I.3.8 [Computer Graphics]: Applications—Applications; J.3 [Computer Applications]: Life and Medical Sciences—Life and Medical Sciences

1. Introduction

Several diagnostic and treatment procedures require the segmentation of anatomical structures from medical images. This can be either performed manually, semi-automatically, or automatically. In manual segmentation, medical experts inspect 2D imaging slices one-by-one, and delineate structures. As this procedure can be time consuming, automatic methods are preferred, with a lot of effort being invested in algorithm development. Still, automatic algorithms cannot account for all cases, and may perform sub-optimally.

Experts working on automatic segmentation algorithms can relatively easily detect the errors. However, even for them, it is not trivial to understand why or how inaccurate outcomes are produced. Exploring and assessing segmentation errors can provide experts with new knowledge about the performance of their algorithms – for example, helping them to predict the anatomic locations and circumstances under which, errors occur. Moreover, it can aid them in confirming or generating hypotheses about their methods and, in the long term, it can allow them to improve their segmentation results. Even if segmentations cannot be improved, it still remains important to be aware of potential inaccuracies. Disregarding this information might affect diagnosis or treatment, if the latter are based on erroneous segmentation outcomes.

As proof-of-concept, we consider the automatic model-based segmentation of pelvic structures [SBV*13], used as input to radiotherapy (RT) planning for prostate tumors. RT is a therapeutic procedure, where tumors are irradiated with a high dose, while the surrounding healthy tissue is preserved. Planning such a procedure requires the accurate segmentation of the prostate and the organs at risk, to be spared during irradiation. Also in this case, segmentation errors often occur, and need to be explored and assessed. This is not feasible, with existing means of exploration.

Our *contribution* is a visual tool that allows experts working on algorithms for model-based segmentation of pelvic structures, to explore and assess the outcomes and errors of their methods. Our approach incorporates the following two capabilities:

- It supports the exploration and assessment of errors in a cohort of pelvic organ segmentations. These segmentations result from applying the same algorithm to several subjects. With this, experts inspect the general performance of the algorithm.
- It facilitates the detailed exploration and assessment of segmentation errors in the pelvic organs of individual subjects. With this, experts can identify the specific details about the performance of the algorithm, concerning each subject of the cohort.

To the best of our knowledge, there is no other tool with the comprehensive functionality that our work offers. Although we demon-

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

DOI: 10.2312/vcbm.20161287

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

strate our visual tool on a specific case, our methods could be generalized to other applications, and suit other segmentation algorithms.

2. Background: Model-based Segmentation of Pelvic Organs

RT is one of the most common treatments for prostate cancer. Its goal is to maximize the effect of the irradiation on tumors, while minimizing the side effects on adjacent healthy organs [WL15]. To this end, planning is performed. For this, accurate segmentation of the prostate and surrounding organs at risk (rectum, bladder, and seminal vesicles) is required. Figure 1 shows the involved anatomy.

For the segmentation of these organs, automatic model-based methods are often employed [EPW*11, SBV*13]. In our work, we consider the algorithm of Schadewaldt et al. [SBV*13], for the segmentation of pelvic structures in CT images. In this method, structures are considered to have a known general shape. Training data are used, to build probabilistic models that explain the shape variation of each structure. These models are used as prior information, and are positioned in the volume. Then, they are iteratively adapted to the boundaries of the structure of interest [EPW*11], using a combination of rules. These are features, such as gradient magnitude, which have been learned from training data. Different features might be employed for different organs, or parts of these. More details about the algorithm can be found in the papers of Schadewaldt et al. [SBV*13] and Ecabert et al. [EPW*11].

Although the selected segmentation method is robust, it is not always accurate. Yet, the resulting inaccuracies can be detrimental for the RT dose administration to healthy organs, with unwanted side effects [WL15]. Our collaborating experts from Philips Healthcare in Hamburg, working on the segmentation of these pelvic structures need to explore, understand, and assess the segmentation results, as well as their respective inaccuracies. To this end, they generate, using their in-house algorithm [SBV*13], segmentations of four organs – prostate, bladder, rectum, seminal vesicles – and their interfaces, in the form of triangulated meshes. Meshes from different subjects have already a triangle-to-triangle correspondence. Additionally, ground truth for each subject is available from delineations of pathologists. Correspondence between the ground truth and the segmentation outcomes has been established, as described in the paper of Schadewaldt et al [SBV*13].

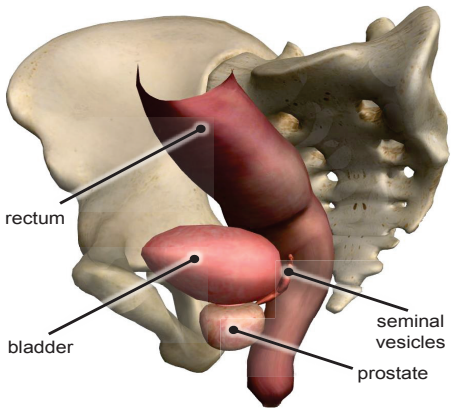

Figure 1: *The anatomy of the pelvic structures involved in this work. Image generated using ZygoteBody™.*

From the ground truth, our segmentation experts computed four local quality measures, per triangle [SBV*13]. These are: **(i)** *target error*, i.e., point-to-point distance from a triangle in the resulting mesh, to the target location in the ground truth data (in mm); **(ii)** *features response*, i.e., the strength of a number of specified algorithm features at the target location; **(iii)** *weighted features response*, i.e., the feature response (inversely) weighted by the distance to the target; and **(iv)** *triangle area* (in mm^2). All these measures are extensively used by our intended users and are indicative of segmentation accuracy. For example, a triangle with high target error is expected to have low feature response, meaning that the selected features are not strong enough to attract the triangle towards the correct target position. Dramatic changes in the triangle areas can be another sign of erroneous segmentations.

Moreover, *feature response profiles* are computed by our collaborating segmentation experts per triangle, after the adaptation of each mesh. As shown in Figure 2, the provided data of the profile of each triangle result into a number discrete point values, along a ray parallel to the normal of the triangle, centered to the adaptation location of each triangle. These values indicate the strength of the features-rules that were used for the adaptation at each position of this ray, and they relate to the above mentioned feature responses and target errors. During adaptation and profile computation, neighboring triangles are influencing each other, as well. For this, profile inspection in triangle neighborhoods, or in groups with similar response, can give a better idea of the reliability, than individual triangles. During such an inspection, the number and locations of peaks, i.e., the local maxima, are important. Multiple peaks could indicate locations with high feature responses that are competing during the adaptation. Non-centered peaks could also be problematic.

After a discussion with our collaborators, it resulted that they currently do not have an intuitive and easy-to-use way, to obtain new insight into their segmentation outcomes, with respect to the computed local quality measures, and the response profiles. They pointed out a number of *tasks* that they are interested in performing:

- For the *full cohort* of subjects:
  - Explore the distribution of local segmentation errors and response profiles **(T1-a)**.
  - Identify anatomical locations (organs or part of these) where the algorithm performs consistently **(T1-b)**.
  - Identify subjects that are special cases **(T1-c)**.
- For an *individual subject*:
  - Explore the distribution and anatomical location of the different local quality measures **(T2-a)**.
  - Discover relations between local quality measures **(T2-b)**.
  - Identify response patterns, for reliability evaluation **(T2-c)**.

3. Related Work

Visual analytics for the exploration of segmentation outcomes has been addressed in several recent papers. Among them, there are frameworks for the *analysis of the impact of parameters* on segmentation algorithm outcomes, such as in the work of Torsney et al. [TWSM*11] and Fröhler et al. [FMH16]. Also, there is recent work on *shape variability analysis* [BBP10, KLR*13, HSK11]. However, the focus of these two paper groups is not on evaluating the employed segmentation algorithms and their results.

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

Table 1: *Requirement analysis concerning our application, for the tasks defined by the intended users and described in Section 2 (✓: fulfilled; x: not fulfilled; ⋆: partially fulfilled, or profile response not fulfilled; grey shading: non-applicable).*

	Compatibility with Data	Multiple Subjects	Compatibility with Tasks					
			(T1-a)	(T1-b)	(T1-c)	(T2-a)	(T2-b)	(T2-c)
Parameter Space Visualizations	✓	x						
Shape Variability Visualizations	x	✓						
Earlier Comparative Visualizations	x	x						
[SPA*14]	x	✓	x	⋆	⋆	x	x	x
[vLAA*13]	x	x	x	x	x	x	x	x
[VLBK*13]	x	✓	x	x	x	x	x	x
[vLBB15]	x	✓	⋆	✓	✓	✓	x	x
[GSK*13]	x	✓	⋆	✓	⋆	✓	⋆	x
Our proposed approach	✓	✓	✓	✓	✓	✓	✓	✓

Another category comprises *comparative visualization*, which rather deals with the direction of qualitative or visual evaluation of two segmentation outcomes, with respect to each other. Busking et al. [BBF*11] proposed visualizations for the comparison of two surfaces, using different kinds of visual or graphical variables. In other papers, simple overlays [GJC01], or extensions of checkerboard visualizations on 2D imaging slices [MHG10, SGB13], but also side-by-side comparisons of 3D volumes have been used [AWH*12]. Visual variables, deformations, glyphs [ZSL*16] and combinations of these have also been employed [PF96]. Specifically for mesh comparison, MeshLab [CCR08] and PolyMeCo [SMS09] have been proposed. Most of these papers refer to comparing two subjects, or one subject with a reference. Comparison of multiple subjects was, only recently, tackled by Schmidt et al. [SPA*14]. In this work, a visual tool for the comparison of meshes is proposed, enabling the interactive exploration of their differences. This tool is meant for evaluating meshes generated by different algorithms, with respect to a reference mesh and it is not fully applicable to the data that we are dealing with. It does not allow to explore and compare any local quality measures along with response profiles, which are necessary for our application. Additionally, it is limited to evaluating the visual quality of the resulting shapes. This is predominantly done in user-selected regions, which need to be interactively inspected.

For the *evaluation of the segmentation process and outcome*, von Landesberger et al. [vLAA*13], visualize the progress of quality during the segmentation of one organ. This approach enables the analysis of the segmentation process, but it is limited to one subject. Later, they improve this by proposing a method to show the distribution of quality values globally, and to select cases with high or low quality values for a detailed inspection [vLBK*13]. This strategy still does not allow the comparison of local quality measures across all subjects. In a more recent paper [vLBB15], von Landesberger et al. present a system for assessing and comparing segmentation quality across multiple datasets. A drill-down approach from an overview of a group of subjects to a detailed view of user-selected cases is employed. As follow-up, Geurts et al. [GSK*15] propose a method for the visual comparison and evaluation of 3D segmentation algorithms. The goal is to determine the best segmentation algorithm, among different alternatives. To this end, they investigated both global and local approaches. Both previous works [vLBB15, GSK*15] are similar to ours, but they are not fully applicable to our available data and tasks. This especially holds for the tasks related to the exploration of the segmentation response and the relations between local quality attributes. Table 1 shows schematically which requirements are (not) fulfilled by the most relevant previous related work.

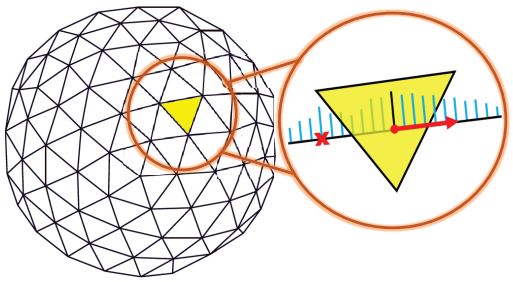

Figure 2: *Response profile (cyan) of a triangle, after mesh adaptation, centered at the adaptation location (black) and parallel to the normal (red). There is also a latent peak, denoted with the cross.*

4. Visual Analytics for the Exploration and Assessment of Segmentation Errors

The segmentation algorithm [SBV*13] is applied on imaging data from a cohort of subjects. Then, the respective triangulated meshes are generated, along with the measures described in Section 2. Our approach enables the exploration and analysis of these measures, using the components shown in Figure 3: **(i)** Exploration of the *full cohort* of subjects (Section 4.1), **(ii)** Exploration of an *error hierarchy*, to detect special subject cases (Section 4.2) and **(iii)** Exploration of an *individual subject* (Section 4.3).

4.1. Exploration of the Full Cohort

When exploring the full cohort of segmentation outcomes, segmentation experts initially need to explore the distribution of local segmentation errors and the respective response profile values **(T1-a)**. Visual comparison of individual outcomes, though, may be time consuming, but also limited, due to perception and screen space constraints [GSK*15]. For this reason, we decided to provide an overview, at a triangle level. As mentioned in Section 2, the individual subjects of the cohort have a triangle-to-triangle correspondence. Thus, at each triangle position, we compute the mean and the standard deviation across all subjects, of both the target error and the response profiles.

For the *target errors*, mean and standard deviation are plotted in

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

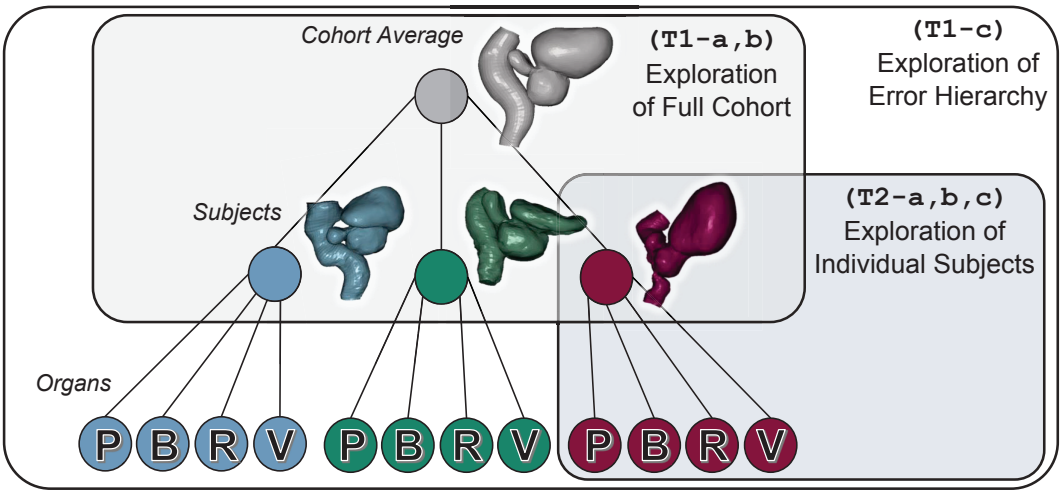

Figure 3: *The three main components of our approach, together with the tasks from Section 2 that they address. The abbreviations denote the different organs (P: prostate, B: bladder, R: rectum, V: seminal vesicles).*

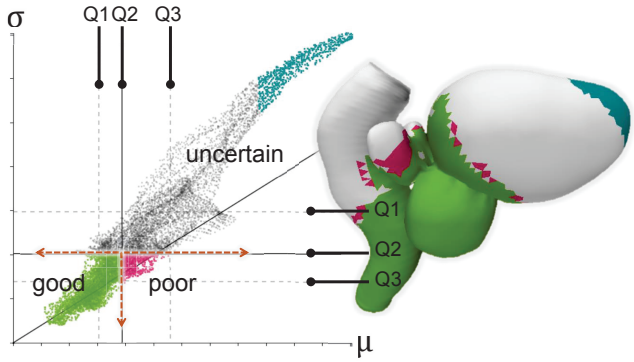

Figure 4: *Confidence scatterplot of the mean error μ against the standard deviation σ, of the target error of all subjects. We denote the three areas of performance (good: ↓μ ↓σ, poor: ↑μ ↓σ, uncertain: ↑σ)* **(T1-a)**. *Three selections are made for good (green), poor (magenta) and uncertain (cyan) performance, and links to the anatomy are shown* **(T1-b)**. *Q1 – Q3 indicate the three quartiles.*

a scatterplot, where each data point represents one triangle location (Figure 4). Data points in the scatterplot are rendered with a lowered opacity, to reduce clutter from overlapping points and as a density indication. We call this representation a *confidence scatterplot*, as it can provide information about three main regions of confidence, based on the values of the mean and the standard deviation of the target error (Figure 4). To convey additional information about the distribution of the mean and the standard deviation of the target error across the triangles of the mesh, we denote the first, second and third quartile of the respective distributions (Figure 4, Q1-Q3). In a confidence scatterplot, points with low mean and low standard deviation represent triangles where the algorithm performs systematically well. Points with high mean and low standard deviation represent triangles where the algorithm performs systematically poorly. Finally, points with high standard deviation correspond to uncertain areas. This is related also to task **(T1-b)**.

For the *response profiles*, a different approach is followed. As already mentioned in Section 2, the reliability of the algorithm can be assessed from the inspection of profiles in triangle neighborhoods – and especially, for triangles with similar response profiles. To this end, the peaks, i.e., the local maxima, are considered. A region of triangles with single-peaked response profile is more likely to be accurate than a region with multiple peaks. One option to illustrate this would be to reduce the mean profile information to a single scalar, representing the number of peaks. However, this would not convey the entire information about the mean profiles. For this, we retrieve clusters of mean profiles with a similar shape. These clusters can, then, be represented and visualized by an average profile.

Several *clustering* approaches can be employed [JMF99]. However, determining a-priori the optimal value of clusters can be difficult and time consuming. For this reason, approaches such as *k*-means were discarded. For our application, we consider it more suitable to use a hierarchical clustering method. The computation of clusters with similar mean profiles is done, using an agglomerative hierarchical clustering method [WJ63]. Initially, the number of clusters is equal to the number of triangles. This is followed by a phase, where iteratively the two most similar clusters are merged. Once a cluster is created, a representative, i.e., average, profile is used in the next iteration. Clustering is performed, with the similarity between two normalized profiles, p and q, being:

$$similarity(p,q) = \sum_i 1 - |p[i] - q[i]| \qquad (1)$$

In this way, two mean profiles with close-by peaks are assigned a higher similarity score, than two mean profiles with peaks further apart. After all iterations are finished, this algorithm results in a dendrogram, which can be interactively browsed. For visualization purposes, we employ a collapsible *profile tree* metaphor, with the root being the average representative profile of all triangles. This can be expanded, revealing all underlying depth levels of clusters. The user can inspect the contents of the clusters interactively, without requiring to define a-priori the preferred number of clusters. Each representative profile from a cluster is depicted in a one-dimensional visualization, shown in Figure 5. In this visualization, the 21 values of the representative profile are normalized to the range $[0..1]$. Each value corresponds to one square and is mapped

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

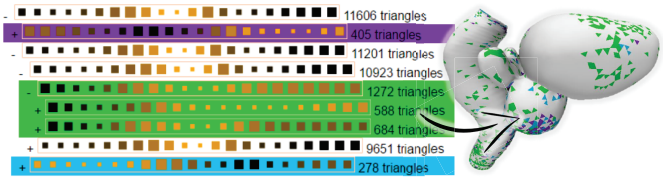

Figure 5: *Profile tree visualization for the exploration of the clustering of profile responses* **(T1–a)**. *The mean (encoded with the color of the squares) and the standard deviation (encoded with the size of the squares) are depicted for the 21 values (squares) of each representative profile (row). Three selections (purple, green, blue) are made to show the link to the anatomy* **(T1–b)**.

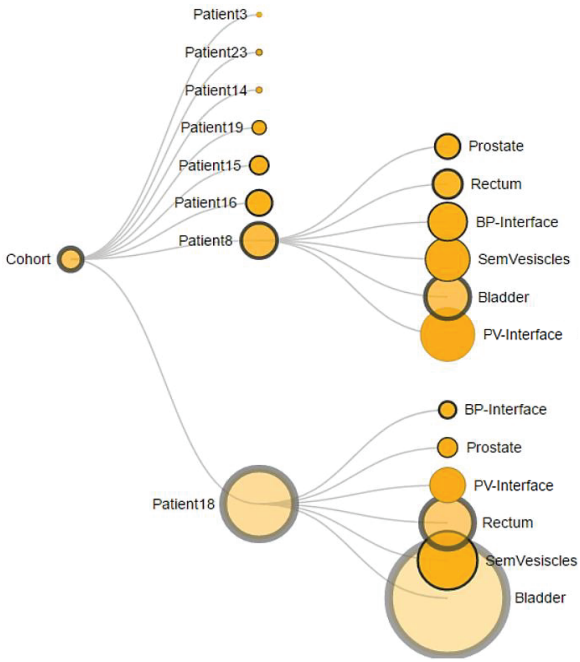

Figure 6: *Exploration of the error hierarchy in the cohort, using a collapsible tree graph representation* **(T1–c)**.

linearly to a single hue, sequential color scale. In Figure 5, peaks are depicted in bright orange, while black denotes local minima. The size of the squares is inversely related to the standard deviation of a representative profile at each of the 21 positions, i.e., smaller squares, indicate larger standard deviations. This encoding was inspired by the work of Höllt et al. [HPvU*16].

After the exploration of errors and profile responses, segmentation experts need to identify whether their algorithm presents coherent behavior. For example, they need to identify the anatomic locations where their algorithm systematically fails or succeeds, at a voxel level **(T1–b)**. To this end, we enable *brushing and linking* both in the confidence scatterplot, and directly on the average reference mesh of the cohort. In this way, we establish a link between the anatomy and the computed target errors (Figure 4). Brushing and linking is applied, also, from the profile tree to the scatterplots (Figure 5). Also, selections in the confidence scatterplot are followed by visualizing the respective average profile. In this way, all components are linked.

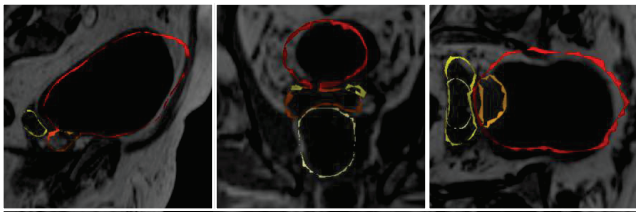

Figure 7: *Qualitative exploration of the intersection of the segmented mesh with the imaging slice data* **(T2–a)** *(red: bladder, orange: prostate, yellow: seminal vesicles, white: rectum).*

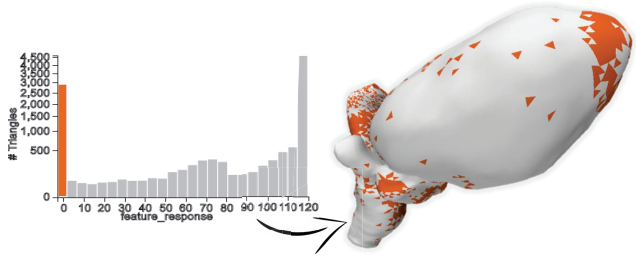

Figure 8: *Interactive exploration of the distribution of local quality measures* **(T2–a)**. *Selections in the histograms provide a link to the anatomy (orange).*

4.2. Exploration of the Error Hierarchy

The next required step is to provide provide an overview on the hierarchy of errors in the full cohort, and allow segmentation experts to identify subjects that are special, interesting cases **(T1–c)**. For this, we employ a collapsible, undirected *tree graph*, to show an overview of the average target error and standard deviation in the full cohort (Figure 6). The root of the tree represents the full cohort, which can be expanded to display the different subjects. These can be further expanded to depict the different organs. The size of the node encodes the magnitude of the average target error, while the magnitude of the standard deviation of the target error is encoded in the opacity of the node, but also in a halo around the circumference (Figure 6). To increase legibility, the nodes are sorted based on the average target error, at each depth level. Also, nodes of the tree that are not interesting for the analysis can be interactively collapsed, to save screen space. This representation summarizes the distribution of target errors in the cohort, across all patients and their respective organs. From this, users can be guided to select individual subjects that need further exploration, in the next stage.

4.3. Exploration of Individual Subjects

Our tool answers also the requirement for a detailed exploration of segmentation errors in individual subjects. The first step involves a qualitative exploration of the resulting segmentation, with respect to the *imaging slice data* (Figure 7). This exploration can give an initial indication of the outcome of the segmentation, as it shows the intersection of the resulting mesh with the imaging data. For the exploration of the distribution and anatomical locations of the different local quality measures **(T2–a)**, *histograms* are employed (Figure 8). Here, interactive selections provide a link to anatomy.

Discovering relations between local quality measures is also necessary **(T2–b)**. For this, we initially enable the pairwise inspection of two measures, directly on the *mesh surface*. This is done

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

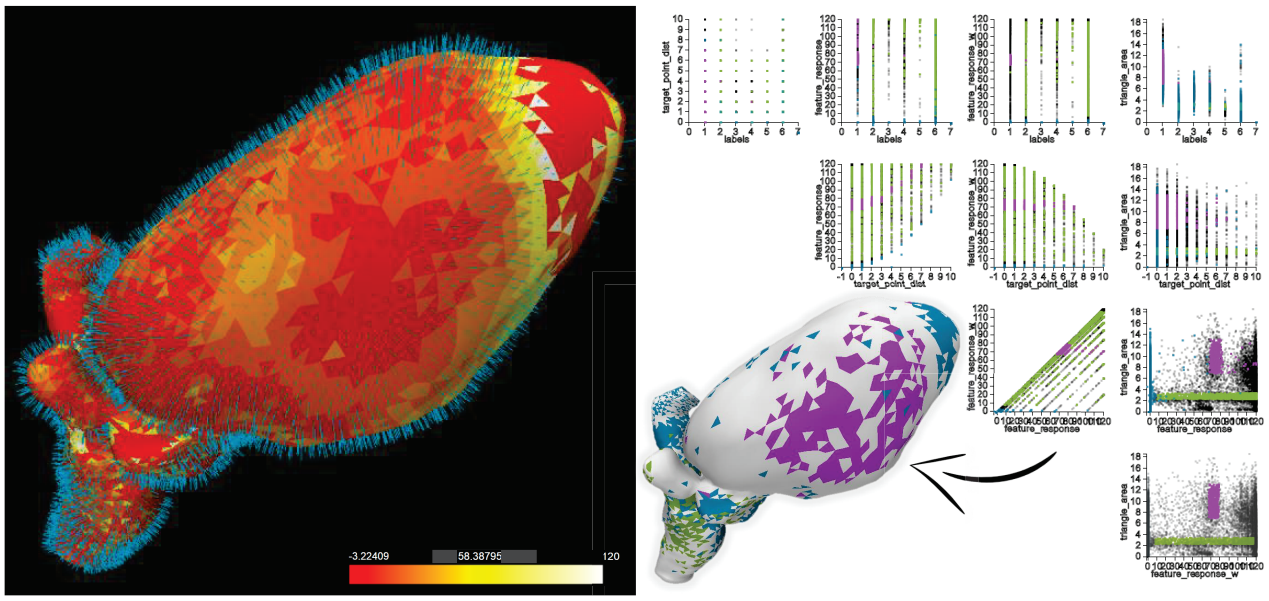

Figure 9: *Discovering the relations between local quality measures (left: comparison of two measures, with color encoding and glyphs, directly on the mesh surface; right: multi-dimensional visualization of local quality measures in a scatterplot matrix, where selections (blue, purple and green) provide a link between different scatterplots, and also to the anatomy)* **(T2-b)**.

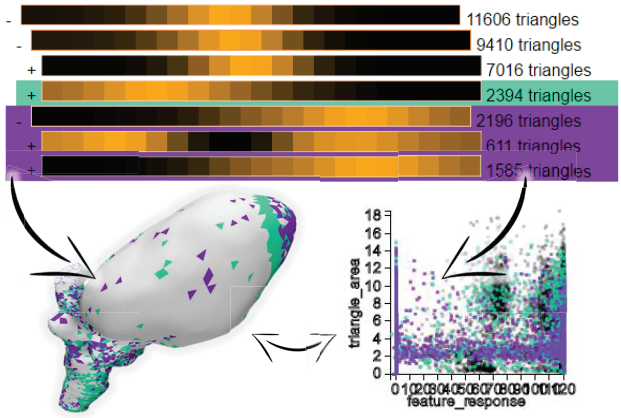

Figure 10: *Profile tree visualization for the exploration of the clustering of profile responses of an individual subject* **(T2-c)**. *Selections (cyan, purple) enable a link to the anatomy and the local quality measures.*

by color encoding one local quality measure with a reduced heated body colormap, and a second measure with line glyphs, along the normal of each triangle of the mesh (Figure 9-left). The size of the glyphs encodes the magnitude of the measure, at each triangle position. Still, this representation limits the exploration of relations to only two dimensions, while glyphs may introduce occlusion. To overcome this, a *scatterplot matrix (SPLOM)* is employed (Figure 9-right). The SPLOM was preferred over other multi-dimensional representations, e.g., parallel coordinate plots, due to the previous familiarity of the intended users. Brushing and linking in the scatterplot matrix facilitates finding and analyzing relations and patterns, across multiple quality measures.

Finally, the identification of patterns in the algorithm response enables segmentation experts to evaluate the algorithm reliability, for each individual subject **(T2-c)**. To this end, we use the same approach, as the one proposed for the cohort exploration. Initially, we retrieve clusters of profiles with similar behavior, using the same hierarchical clustering method, as in task **(T1-c)**. Then, a similar *profile tree* metaphor is employed. Here, each representative profile is depicted in a one-dimensional visualization that highlights the peaks of the profile clusters (Figure 10). As we have only one subject, the standard deviation encoding is not necessary. Interaction is employed to enable browsing the clustering hierarchy. Also, if a cluster is selected in the profile tree, the respective quality measures and the anatomical location are highlighted, in the SPLOM and the mesh, respectively (Figure 10).

Implementation. The application is developed in WebGL, using Three.js and D3.js. It is compatible with all browsers and platforms.

5. Results: Usage Scenario

In this section, we elaborate on a usage scenario. Our purpose is to illustrate the functionality and some initial results that can be achieved with our proposed visual tool. This usage scenario has been guided by our collaborating segmentation experts, based on their previous knowledge and expectations. It was used to *explore* their data and to *confirm hypotheses* about their algorithm.

5.1. Dataset

The dataset employed for this usage scenario consists of a cohort of eight subjects. The explored data consisted of: **(i)** CT volumetric data for all eight subjects, with dimensions $320 \times 320 \times 120$ and a spatial resolution of $1.563 \times 1.563 \times 1$ mm. **(ii)** A reference (average) mesh and the meshes of the eight subjects, each containing $11,606$ triangles with a triangle-to-triangle correspondence and organ labels. **(iii)** The respective local quality measures per-triangle. **(iv)** The 21-valued profile response data, per-triangle.

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

5.2. Exploration of the Full Cohort

For the exploration of the full cohort, the average reference mesh is employed, together with the mean and standard deviation of the target error, and the mean and the standard deviation of the profile responses, per-triangle. Initially, the distribution of the local segmentation errors and profiles **(T1–a)**, and also their anatomical correspondence **(T1–b)** are explored. In Figure 4, we illustrate in the confidence scatterplot the mean target error against the standard deviation at each triangle position, for the full cohort. From this, we can divide the algorithm performance into three categories: good, poor and uncertain. The good (green) and poor (magenta) categories are much less dispersed than the uncertain one.

Through brushing and linking, we can identify the anatomic regions of good performance, which correspond to the prostate and also its very adjacent surfaces (Figure 4, green). These are the parts, where the algorithm achieves high precision and high accuracy. Poor performance can be seen mainly in the seminal vesicles (Figure 4, magenta), which can be explained by the fact that seminal vesicles are small structures that may be hard to discern, and also are highly variable in shape. These are the parts, where the algorithm achieves low accuracy, but high precision. The rest, i.e., the biggest part of the bladder and also the top half of the rectum belong to the uncertain performance category. In particular, triangles of the bladder or the rectum that are further away from the prostate (cyan) are more uncertain, i.e., have a low accuracy and low precision. This might be related to the high variability in the shape of these two organs. An additional reason for this might be that the employed algorithm produces segmentations used for RT planning in patients with prostate tumors. This is expected to affect structures closer to the prostate. Thus, the segmentation algorithm might be promoting better results for parts closer to the prostate.

The profile exploration in Figure 5 shows interesting results, as well. Despite the fact that some triangles have an unusual profile response, i.e., a profile where the peak was not in the middle (Figure 5, purple, green and blue) these triangles still manage to achieve a low target error, as they are influenced by neighboring triangles.

5.3. Exploration of Error Hierarchy

By exploring the error hierarchy in the tree graph **(T1–c)**, we can identify the subjects and organs, where larger errors appear. From the representation illustrated in Figure 6, we identify *Patient18* as the subject with larger errors, and *Patient3* as the subject with smaller errors. For *Patient18*, the segmentation of the bladder has the largest error, while the interface between the bladder and the prostate seems to be well-segmented. For *Patient3*, the segmentation of all organs and their interfaces has small errors. *Patient8* is also another interesting case, where most of the organs have a high error. Here, the bladder and the interface between the prostate and the seminal vesicles segmentations have the highest errors, as depicted in Figure 6. From this exploration, we can select which subjects need to be explored individually, in more detail.

5.4. Exploration of Individual Subjects

We explore individually two cases - *Patient18* and *Patient3* - identified previously, as the worst and best results, respectively.

First case – *Patient18*. For this subject, the segmentation outcome had the largest error. An initial qualitative exploration of the intersection of the segmented mesh, with respect to the imaging slice data **(T2–a)**, as illustrated in Figure 7, indicates the locations where the algorithm had a bad performance. The prostate contour, denoted with white in the coronal slice (middle), seems to be well aligned with the borders of the organ on the CT slice. However, the bladder (red) is not. The tip of the bladder has been missed and also a distal shift is visible. The histograms in Figure 8 **(T2–a)** show a large peak in the distribution of *feature response* measure (second row, left) at the zero value, but also at the maximum value. The first peak indicates that for many triangles no suitable feature could be discovered, and that there may have been a problem with the feature selection. The respective triangles are located at the top of the rectum and on the tip of the bladder (Figure 8, orange). The second peak corresponded to well-segmented triangles.

Upon inspection of the *feature response* against the *triangle area* measure **(T2–b)** in Figure 9-left, we see that the tip of the bladder corresponds to a low feature response, denoted with the red color, and to low triangle area values, denoted with smaller-sized glyphs. We confirm this also in the SPLOM (Figure 9-right). Several clusters are easily identified when plotting the two measures, in the scatterplot in the fourth column and third row. These clusters include one with low response values (blue), one with low triangle area (green) and one cluster in the middle (purple). Selections provide insight into the physical location of these clusters (Figure 9-right), revealing interesting information. The purple cluster corresponds to the areas at the sides of the bladder that presented the distal shift, in the previous qualitative exploration. The blue cluster corresponds to the wrongly segmented tip of the bladder and top of the rectum, and the green cluster corresponds to the well segmented regions of the bottom part of the rectum and the prostate. This exploration suggests a lack of strong features in the bladder and top of the rectum that may be responsible for the errors.

The profile information of this subject is also investigated **(T2–c)**. Several triangles highlighted in Figure 10 are triangles where the feature response profile did not contain any peak close to the middle (cyan, purple). In practice, it is acceptable if a triangle has a feature response profile without peaks, as long as most of the neighbors do not present this same behavior. The position of neighboring triangles can positively influence the position of a triangle. In our case, these cyan and purple triangles add up to a total of 4,590, which is almost $\frac{1}{3}$ of the mesh and they seem to form in their majority coherent regions. Therefore, an absence of peaks in these profiles indicates that no information was available on how to modify the triangles location and that the current location of the triangle is not supported by any of the selected features.

Second case – *Patient3*. For this subject, the segmentation outcome had the smallest error. An initial qualitative exploration of the intersection of the segmented mesh with respect to the imaging data **(T2–a)** showed that the segmentation outcome matches well the borders of the organs in the slices (Figure 11-a). The histograms **(T2–a)** show a large peak in the distribution of the *feature response* measure at the zero value, but also at the maximum value (second row, left). The first peak indicates that for some triangles no suitable feature could be discovered, while the second peak corresponded to well-segmented triangles with a high feature response

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

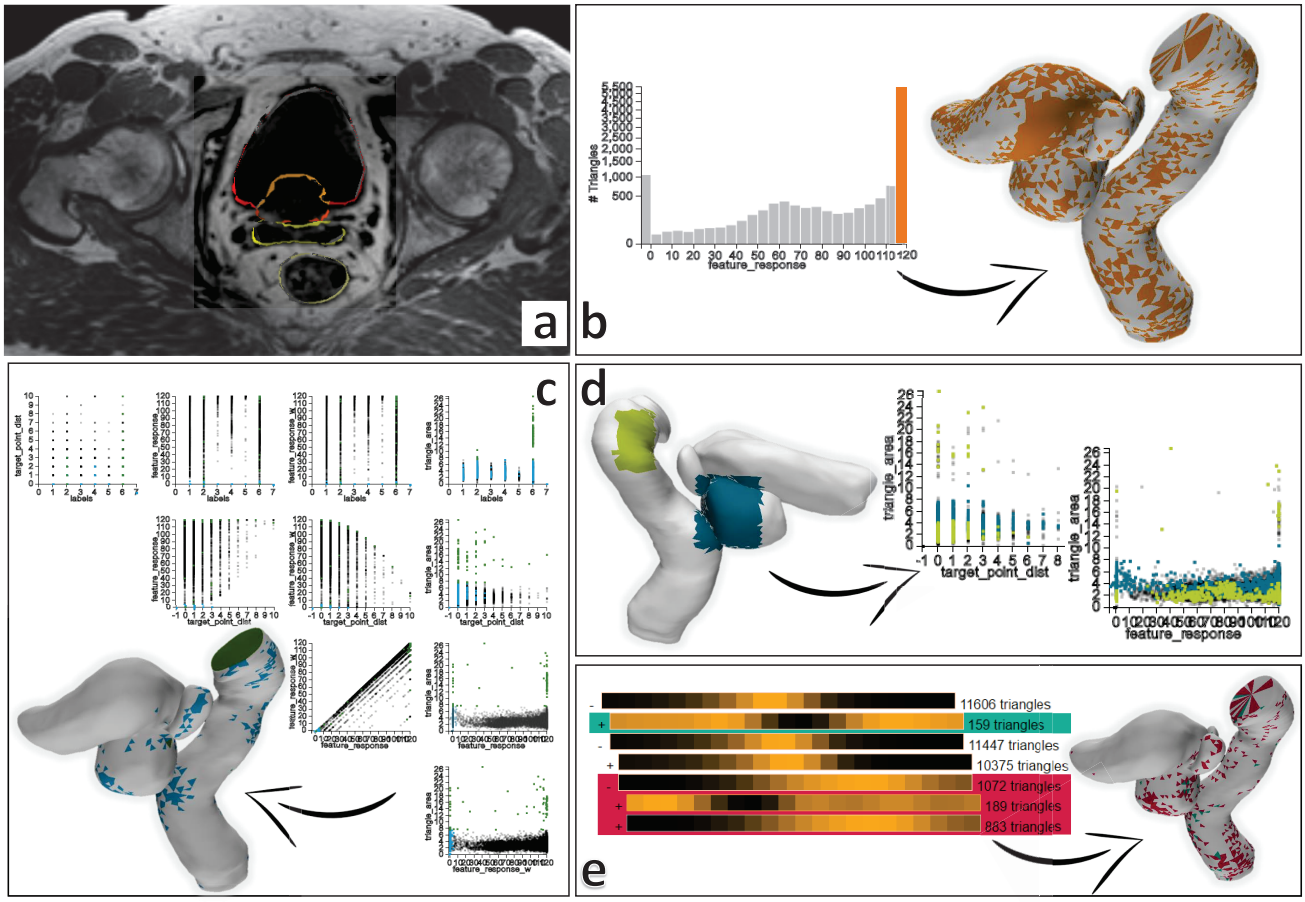

Figure 11: *Usage scenario for a subject, where the algorithm has a good performance (Patient3). (a) Exploration of the intersection of the mesh with imaging slice data; (b) Exploration of the distributions of the quality measures and link to the anatomy (orange); (c) Exploration of relations between quality measures and link to the anatomy; (d) Selections directly on the mesh (green and blue), and exploration of corresponding quality measures; (e) Profile response exploration, and link to the anatomy.*

(Figure 11-b, orange). In the SPLOM **(T2-b)**, the majority of points has a low triangle area, but there are several data points, i.e., triangles, with an outlier behavior (Figure 11-c, green). Upon selecting those data points for further exploration, they correspond to the triangles on the top and the bottom of the rectum. We are also interested in seeing which parts of the meshes correspond to a low feature response (Figure 11-c, blue). These parts are few and scattered around the mesh. They have mostly a target error smaller than 4*mm* (second row, fourth column).

Another approach to investigate the segmentation quality is to select triangles on the mesh and inspect the attribute value distributions in the scatterplots. Figure 11-d shows a selection containing the prostate and a small part of the adjacent organs and, also, a selection far away from the prostate, on the rectum (green). The scatterplots show the distribution of the selected triangles mapping the feature response against the triangle area and the target point distance. The majority of triangles far from the prostate (blue) have both high and low triangle area. The selection on the rectum (green) has low triangle area. As mentioned before, dramatic changes in these values indicate segmentation errors. The profile information of this subject is also investigated **(T2-c)**. Only few triangles (1,231) have a profile without a peak in the middle. These are almost $\frac{1}{10}$ of all triangles, and they are evenly spread through the whole mesh (Figure 11-e, cyan and magenta).

6. Evaluation

To assess the value of our visual tool, we designed an evaluation, inspired by the paper of Lam et al. [LBI*12]. The evaluation was performed with five experts, working on developing segmentation algorithms. The group of participants included one professor in the field of Medical Image Analysis, three research scientists in the field of Image Processing and one scientist in the field of Computer Science. We did not include clinical experts, as they are not the intended users of our tool. Their experience with segmentation algorithms varies from seven years (for two people) to more than twenty years (for one person). All of them have also a radiological background. Four evaluators are male and one is female. They all have normal vision, two wear glasses and nobody is colorblind.

The evaluation had to be conducted remotely, and the participants were not able to interact with the tool. During the session, we demonstrated step-by-step the visual tool, using data provided by the experts and well-known to them. We demonstrated the main

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

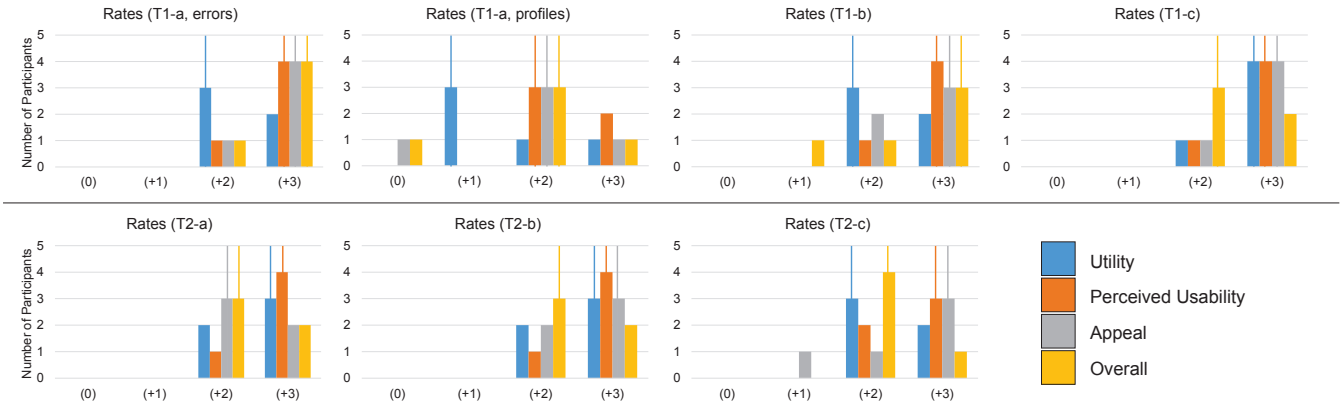

Figure 12: *Rating results for the first part of our evaluation, for each of the tasks of our proposed tool. The scale range is* $[-3..+3]$*, but we only received answers higher than* 0*. With the additional vertical lines, we denote the median of each rating.*

components of the tool, simulating the visual environment for the exploration and analysis of segmentation errors. The evaluation participants followed the demonstration. They could interrupt at any moment to make exploratory requests, e.g., selections and interactions that could help them analyze and understand their observations. We allowed them to discuss with each other these observations, but not their opinion on the tool.

To document their opinion on the demonstrated visual tool, they completed a questionnaire. This consisted of two parts. The questions of the first part were related to the tasks, presented in Section 2. We divided task **(T1-a)**, to evaluate separately the visualizations employed for the error distributions and for the profile responses. Each question required an open answer, and also rating in a seven-point scale $[-3..+3]$. We evaluated four aspects: Utility (*Does it do what it is meant to do?*), Perceived Usability (*Would I be able to learn and use it?*), Appeal (*Do I like it?*) and Overall Feeling (*How do I feel about it, in general?*). The second part of the questionnaire included several questions regarding strengths, limitations, missing features of the tool and proposals for improvement.

Ratings. Figure 12 summarizes the results of the first part of our evaluation. There was no correlation with respect to the experience level of the evaluation participants. Most of the evaluated aspects ranked on the positive side of the scale (≥ 1), while only two received a neutral grade (0). All aspects have a median value of at least two, apart from one that has a median of one. The lower values were all documented for the profile response part of task **(T1-a, profiles)** and were all given by the same person. The error distribution part of task **(T1-a, errors)** and the error hierarchy exploration **(T1-c)** were, in general, rated higher than the rest.

Open Answers. The above mentioned ratings are in agreement with the open answers of the first part, and also with the second part of the questionnaire. The evaluators considered the tool to be overall intuitive and potentially easy to use. One evaluator commented that "*it is a light-weight web-based tool, which makes it highly optimized for model-based segmentation analysis*", due to the involved large data. The feature that received most positive comments was the dynamic selection of triangles on the meshes, on the scatterplot and also their in-between link, i.e., **(T1-a, errors)**, **(T1-b)**, **(T2-a)** and **(T2-b)**. Yet, for the selection on the mesh, an evaluator commented that he would like to have visual feedback for the selections. Another appreciated feature was the tree graph, for exploring the error hierarchy in the cohort **(T1-c)**. One evaluator commented that he would actually like to use it, to explore a much larger cohort of segmentation outcomes.

Feedback for Improvement. Most of the participants gave feedback about improving the cohort profile response part **(T1-a, profiles)**. First, they commented that the visualization of the cohort profile responses takes some time to understand. It does not allow to change the similarity measure, apart from the mean values during clustering. One participant commented that the representation for the profiles in the cohort can be even reworked to be presented as an average curve, with a confidence band that denotes variability. This is in contrast to the positive opinion that the evaluators expressed for the individual profile response part **(T2-c)**, which was considered more intuitive and rated much higher. Another participant disliked, in particular, the glyphs used in task **(T2-b)**. These limitations were proposed as points for future work, along with a simultaneous visualization of multiple profile data. Also, functionality for annotating observations and for making a report from these, along with captured screenshots was proposed as future work.

7. Conclusions and Future Work

We introduced a visual tool to enable experts, working on algorithms for the model-based segmentation of pelvic organs, to explore and analyze the outcomes and errors of their methods. Our approach supports the global exploration of errors in a cohort of pelvic organ segmentations, where the performance of the algorithm can be assessed. Also, it enables the exploration and assessment of segmentation errors for individual subjects. We demonstrated the functionality of our tool with a usage scenario. Also, we performed an initial evaluation with five segmentation algorithm researchers, who confirmed the exploratory value of the tool, and gave feedback for future improvements.

A direction for future work includes improving the functionality for the exploration of the profile responses in the cohort. Adding functionality for the comparison of different aspects of the data, such as the local quality errors and profiles of different subjects, is also important. Exploring the impact of parameters used in the segmentation, and also the relation of the shapes of the various organs

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.

to the algorithm performance would be another interesting enrichment. An additional evaluation to quantify the experience of the user is needed, and will be conducted in the future. The proposed visual tool is a promising basis for segmentation experts. It allows them to gain more knowledge on the performance of their segmentation algorithms, and to determine strategies to improve their segmentation results.

8. Acknowledgements

This work was supported by the FP7 European Project DR THERAPAT - Digital Radiation Therapy Patient. We thank Philips Research Hamburg - in particular, Nicole Schadewaldt and Steffen Reinisch - for their data, extensive collaboration and feedback. We also thank all the evaluation participants.

References

[AWH*12] ALABI O. S., WU X., HARTER J. M., PHADKE M., PINTO L., PETERSEN H., BASS S., KEIFER M., ZHONG S., HEALEY C.: Comparative visualization of ensembles using ensemble surface slicing. In *IS&T/SPIE Electronic Imaging* (2012), International Society for Optics and Photonics, pp. 82940U–82940U. 3

[BBF*11] BUSKING S., BOTHA C. P., FERRARINI L., MILLES J., POST F. H.: Image based rendering of intersecting surfaces for dynamic comparative visualization. *The visual computer 27*, 5 (2011), 347–363. 3

[BBP10] BUSKING S., BOTHA C. P., POST F. H.: Dynamic Multi-View Exploration of Shape Spaces. In *Computer Graphics Forum* (2010), vol. 29, Wiley Online Library, pp. 973–982. 2

[CCR08] CIGNONI P., CORSINI M., RANZUGLIA G.: MeshLab: An open-source 3D mesh processing system. *Ercim news 73*, 45-46 (2008), 6. 3

[EPW*11] ECABERT O., PETERS J., WALKER M. J., IVANC T., LORENZ C., VON BERG J., LESSICK J., VEMBAR M., WEESE J.: Segmentation of the heart and great vessels in CT images using a model-based adaptation framework. *Medical image analysis 15*, 6 (2011), 863–876. 2

[FMH16] FRÖHLER B., MÖLLER T., HEINZL C.: GEMSe: Visualization-Guided Exploration of Multi-channel Segmentation Algorithms. *Computer Graphics Forum (Proceedings of the Eurographics Conference on Visualization (EuroVis) 2016 35*, 3 (June 2016). 2

[GJC01] GERIG G., JOMIER M., CHAKOS M.: Valmet: A new validation tool for assessing and improving 3D object segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001* (2001), Springer, pp. 516–523. 3

[GSK*15] GEURTS A., SAKAS G., KUIJPER A., BECKER M., VON LANDESBERGER T.: Visual comparison of 3D medical image segmentation algorithms based on statistical shape models. In *Digital Human Modeling. Applications in Health, Safety, Ergonomics and Risk Management: Ergonomics and Health*. Springer, 2015, pp. 336–344. 3

[HPvU*16] HÖLLT T., PEZZOTTI N., VAN UNEN V., KONING F., EISEMANN E., LELIEVELDT B. P., VILANOVA A.: Cytosplore: Interactive Immune Cell Phenotyping for Large Single-Cell Datasets. *Computer Graphics Forum 35*, 3 (2016), 171–180. 5

[HSK11] HERMANN M., SCHUNKE A. C., KLEIN R.: Semantically steered visual analysis of highly detailed morphometric shape spaces. In *Biological Data Visualization (BioVis), 2011 IEEE Symposium on* (2011), IEEE, pp. 151–158. 2

[JMF99] JAIN A. K., MURTY M. N., FLYNN P. J.: Data clustering: a review. *ACM computing surveys (CSUR) 31*, 3 (1999), 264–323. 4

[KLR*13] KLEMM P., LAWONN K., RAK M., PREIM B., TÖNNIES K. D., HEGENSCHEID K., VOLZKE H., OELTZE S.: Visualization and analysis of lumbar spine canal variability in cohort study data. In *VMV* (2013), pp. 121–128. 2

[LBI*12] LAM H., BERTINI E., ISENBERG P., PLAISANT C., CARPENDALE S.: Empirical studies in information visualization: Seven scenarios. *Visualization and Computer Graphics, IEEE Transactions on 18*, 9 (2012), 1520–1536. 8

[MHG10] MALIK M. M., HEINZL C., GROELLER M. E.: Comparative visualization for parameter studies of dataset series. *Visualization and Computer Graphics, IEEE Transactions on 16*, 5 (2010), 829–840. 3

[PF96] PANG A., FREEMAN A.: Methods for comparing 3D surface attributes. In *Electronic Imaging: Science & Technology* (1996), International Society for Optics and Photonics, pp. 58–64. 3

[SBV*13] SCHADEWALDT N., BYSTROV D., VIK T., SCHULZ H., PETERS J., FRANZ A., BUERGER C., BZDUSEK K.: Robust initialization of multi-organ shape models. *In MICCAI Challenge Workshop on Segmentation: Algorithms, Theory and Applications* (2013). 1, 2, 3

[SGB13] SCHMIDT J., GRÖLLER M. E., BRUCKNER S.: VAICo: Visual analysis for image comparison. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 2090–2099. 3

[SMS09] SILVA S., MADEIRA J., SANTOS B. S.: PolyMeCo - An integrated environment for polygonal mesh analysis and comparison. *Computers & Graphics 33*, 2 (2009), 181–191. 3

[SPA*14] SCHMIDT J., PREINER R., AUZINGER T., WIMMER M., GRÖLLER M. E., BRUCKNER S.: YMCA – your mesh comparison application. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on* (2014), IEEE, pp. 153–162. 3

[TWSM*11] TORSNEY-WEIR T., SAAD A., MÖLLER T., HEGE H.-C., WEBER B., VERBAVATZ J.-M., BERGNER S.: Tuner: Principled parameter finding for image segmentation algorithms using visual response surface exploration. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 1892–1901. 2

[vLAA*13] VON LANDESBERGER T., ANDRIENKO G., ANDRIENKO N., BREMM S., KIRSCHNER M., WESARG S., KUIJPER A.: Opening up the black box of medical image segmentation with statistical shape models. *The Visual Computer 29*, 9 (2013), 893–905. 3

[vLBB15] VON LANDESBERGER T., BASGIER D., BECKER M.: Comparative Local Quality Assessment for 3D Medical Image Segmentation with Focus on Statistical Shape Model-based Algorithms. *Visualization and Computer Graphics, IEEE Transactions on* (2015). 3

[vLBK*13] VON LANDESBERGER T., BREMM S., KIRSCHNER M., WESARG S., KUIJPER A.: Visual Analytics for model-based medical image segmentation: Opportunities and challenges. *Expert Systems with Applications 40*, 12 (2013), 4934–4943. 3

[WJ63] WARD JR J. H.: Hierarchical grouping to optimize an objective function. *Journal of the American statistical association 58*, 301 (1963), 236–244. 4

[WL15] WASHINGTON C., LEAVER D.: *Principles and Practice of Radiation Therapy*. Elsevier Health Sciences, 2015. 2

[ZSL*16] ZHANG C., SCHULTZ T., LAWONN K., EISEMANN E., VILANOVA A.: Glyph-based comparative visualization for diffusion tensor fields. *Visualization and Computer Graphics, IEEE Transactions on 22*, 1 (2016), 797–806. 3

© 2016 The Author(s)
Eurographics Proceedings © 2016 The Eurographics Association.