INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
2017, VOL. 33, NO. 2, 123–134
http://dx.doi.org/10.1080/10447318.2016.1220729

Taylor & Francis
Taylor & Francis Group

Using GOMS and NASA-TLX to Evaluate Human–Computer Interaction Process in Interactive Segmentation

Anjana Ramkumar[a], Pieter Jan Stappers[a], Wiro J. Niessen[b,c], Sonja Adebahr[d,e], Tanja Schimek-Jasch[d], Ursula Nestle[d,e], and Yu Song[a]

[a]Faculty of Industrial Design Engineering, Delft University of Technology, Delft, The Netherlands; [b]Department of Radiology and Medical Informatics, Erasmus MC, Rotterdam, The Netherlands; [c]Faculty of Applied Sciences, Delft University of Technology, Delft, The Netherlands; [d]Department of Radiation Oncology, University Medical Center, Freiburg, Germany; [e]Department of Radiation Oncology, German Cancer Consortium (DKTK), Heidelberg, Partner Site Freiburg, Germany

ABSTRACT

HCI plays an important role in interactive medical image segmentation. The Goals, Operators, Methods, and Selection rules (GOMS) model and the National Aeronautics and Space Administration Task Load Index (NASA-TLX) questionnaire are different methods that are often used to evaluate the HCI process. In this article, we aim at improving the HCI process of interactive segmentation using both the GOMS model and the NASA-TLX questionnaire to: 1) identify the relations between these two methods and 2) propose HCI design suggestions based on the synthesis of the evaluation results using both methods. For this, we conducted an experiment where three physicians used two interactive segmentation approaches to segment different types of organs at risk for radiotherapy planning. Using the GOMS model, we identified 16 operators and 10 methods. Further analysis discovered strong relations between the use of GOMS operators and the results of the NASA-TLX questionnaire. Finally, HCI design issues were identified, and suggestions were proposed based on the evaluation results and the identified relations.

1. Introduction

Segmentation is an intermediate step in the analysis of images where regions of interest (ROI) are isolated from the background in order to make the representation of a volumetric image stack more meaningful and easier for subsequent analysis or visualization (Olabarriaga & Smeulders, 2001). In healthcare, segmentation of medical images is needed to support tasks such as diagnosis, prognosis, and planning of medical interventions. For instance, in radiotherapy planning, accurate segmentations of tumors and organs at risk (OAR) are prerequisites for maximizing the delivery of radiation dose to the tumor while sparing the normal tissues in the treatment (Ramkumar et al., 2015). In the segmentation of OAR, a stack of 2D medical images, usually Computational Tomography (CT), Magnetic resonance images (MRI), and/or Positron Emission Tomography (PET) images which shows anatomical and/or physiological information of the subject, is presented to the physician. He/she segments OAR using a type (or a combination) of the manual, semi-automatic or automatic segmentation methods. Figure 1 shows an example of the image segmentation of the heart (a type of OAR). At the left, a 2D CT image is presented to the physician in the axial direction and he/she can draw/create the contour of the heart using either of the segmentation methods. At the right, contours on each 2D image are aligned in 3D and later they can be interpolated to a 3D volume.

With the increasing amount of imaging data acquired during a scanning, automated segmentation methods have attracted much attention in the past decade (Balafar, Ramli, Saripan, & Mashohor, 2010; Petitjean & Dacher, 2010; De Boer et al., 2010). However, physicians' expertise to combine observed image data with prior clinical knowledge to accurately perform segmentation is still the rule rather than exception. Manual contours are in most cases still considered to be the reference standards by many researchers (Hammers et al., 2007; Von Falck et al., 2010), however the process are often time consuming and the results are prone to inter- and intra-observer variabilities. Interactive segmentation methods, which use physicians' expertise to guide the data-driven automatic algorithms in the segmentation process, have been developed in many research projects and commercial software (Olabarriaga & Smeulders, 2001; McGuinness & O'Connor, 2010; McGuinness & O'Connor, 2011; Heckel et al., 2013). In an interactive segmentation process, physicians are asked to give inputs either in the pre- or post-processing or during the segmentation process, depending on the algorithm(s) and the designed workflow. Previous research indicates that besides computational algorithm(s), the efficiency of the interactive segmentation also highly depends on the design of Human–Computer Interactions (HCI, Olabarriaga & Smeulders, 2001, Ramkumar et al., 2015).

In an interactive segmentation process, various HCI components play an important role such as: 1) user input devices (UIDs); 2) user input tools; and 3) types of user inputs. Mouse, keyboard,

CONTACT Yu Song y.song@tudelft.nl Faculty of Industrial Design Engineering, Landbergstaraat 15, Delft University of Technology, 2628CE, Delft, The Netherlands.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hihc.

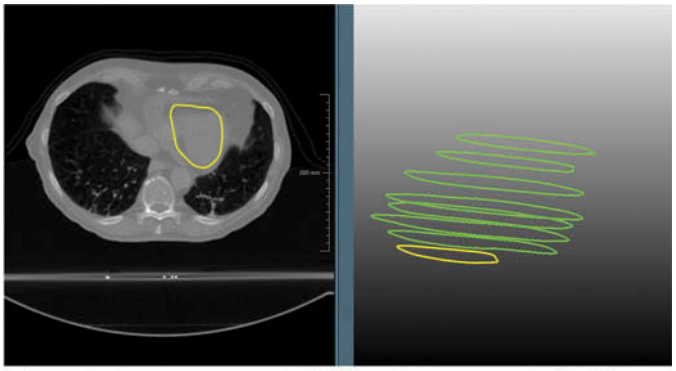


Figure 1. An interface showing segmentation of the heart on CT images.

and monitor screen are often used as input devices to achieve desired HCI. In addition, many other devices may facilitate this process as well. For instance, Harders and Székely (2003) evaluated the value of haptic feedback in a multimodal image segmentation task and found that the used approach is only applicable to linear structures. Lundström, Rydell, Forsell, Persson, and Ynnerman (2011) discovered that a touch interface was very intuitive during pre-operative planning and also helpful in getting: 1) a better understanding of complex anatomy, 2) a better support for planning the surgery procedure, and 3) a better foundation for follow-up assessments. A study conducted by Kotani and Horii (2003) revealed that muscular load of using a mouse exceeded that of using a pen. Sherbondy et al. (2005) evaluated the use of a trackball, pen-tablet, and mouse for segmentation. They found that the pen-tablet UID in two distinct configurations performed faster than the mouse and trackball UIDs in a simulated angiography localization task. Murata (2006) discovered that an eye gaze input system led to a faster pointing time as compared with mouse input, especially for older adults.

With various input devices, users may select different input tools to perform interactions. Olabarriaga and Smeulders (2001) investigated HCI issues in 2D segmentation tasks and found that deforming contours and editing boundaries were the most frequent tools used in segmentation. Aselmaa et al. (2013) discovered that in manual segmentation, brush tool, 3D pencil, smart brush, and nudging tools were frequently used. Kang, Engelke, and Kalender (2004) developed three types of editing tools: hole-filling, bridging points using lines to form a contour and surface-dragging. They concluded that the efficiency of interactive segmentation may be improved significantly by including 3D editing tools in the early stage of the design process.

Using HCI tools, users may provide different types of inputs: for instance, a user may make a series of clicks around target boundaries (Mortensen & Barrett, 1998) or draw sample regions (Boykov & Jolly, 2001; Karasev et al., 2013) to guide the segmentation process. The intuitiveness of the tools is critical in designing a useful interactive segmentation method. Zhao & Xie (2013) classified the user inputs used in segmenting medical images into three categories: menu option selection, pictorial input on an image grid, and parameter tuning. Among those three types of user inputs, menu option selection is considered as the most efficient way, but it only offers limited choices to the user. Pictorial input is simple, but it could be time-consuming. For instance, the user has to draw a contour precisely on an image grid. Tuning parameters of the computational parameters is an easy operation, but it may require specific training for insights of the computational algorithm to select the correct parameters. Yang, Cai, Zheng, and Luo (2010) concluded that the type of user input is an important factor for design interactive segmentation methods as it also affects the efficiency of the process and the outcome of the segmentation results.

In this article, we aim at improving the HCI process of interactive segmentation methods for radiotherapy. Through a case study and using two different types of HCI inputs, we evaluated the HCI process using both an analytical GOMS model and a subjective NASA-TLX questionnaire. Our objectives are to: 1) identify the relations between the GOMS model and the NASA-TLX questionnaire to get a better understanding of analytical and subjective measures and 2) combine those findings to evaluate the HCI process in the interactive segmentation in order to propose HCI design suggestions.

The remainder of this article is organized as follows: Section 2 reviews different types of evaluation methods that have been applied in HCI evaluation. In Section 3, the prototypes which will be used in the experiment are described. The experimental setup is shown in Section 4 and experimental results are presented and analyzed in Section 5. The findings are discussed in Section 6 where suggestions for the HCI design are presented as well. Finally, a short conclusion is drawn in Section 7.

2. Review of HCI Evaluation Methods

A variety of evaluation methods have been used in the literature to assess the HCI process. Gao, Wang, Song, Li, and Dong (2015) classified HCI evaluation measures into four categories: subjective measures, performance measures, psychophysiological measures, and analytical measures. In this article, we focused on subjective measures and analytical measures, as the performance measure is more about the accuracy than the HCI process and many psychophysiological measures are intrusive, which may influence the behavior of the user (Dirican & Göktürk, 2011).

Subjective measures are designed to collect the opinions from the operators about the workload/human effort, satisfaction, preference, user-experience, etc. In spite of the criticism on the validity and vulnerability to personal bias of those self-report methods, subjective measures with the low cost and ease of administration, as well as adaptability, have been found highly useful in a variety of domains, including healthcare, aviation, driving and even office working environment (Bridger & Brasher, 2011; Longo & Kane, 2011; Chang, Hwang, & Ji, 2011; Morgan & Hancock, 2011; Roscoe & Ellis, 1990). The most common way of obtaining subjective measure is through questionnaires. The National Aeronautics and Space Administration Task Load Index (NASA-TLX, Hart & Staveland, 1988) questionnaire is one of the most widely used instruments and has been extensively tested in human factors studies for the measurement of workload. NASA-TLX consists of a set of six rating scales to evaluate the workload of the users in a task (Hart & Staveland, 1988). Those six rating scales are mental demand, physical demand, temporal demand, performance, effort and frustration. Each rating scale is divided into 21 gradations. An example of the NASA-TLX

questionnaire is presented in Appendix 1. The comparisons of sensitivity and diagnosticity between NASA-TLX questionnaire and other subjective measures have been a long and on-going debate, but NASA-TLX questionnaire consistently exhibits high reliability, user acceptance and low inter-subject variability in various studies (Cain, 2007; Dey & Mann, 2010; Rubio, Diaz, Martin, & Puente, 2004). NASA-TLX was used in HCI studies to identify users' emotions, metal demands (Jeon & Croschere, 2015), performance (Gao et al., 2015), etc. In radiotherapy software design, several studies have been using the NASA-TLX questionnaire to identify physicians' workload during various stages of the workflow (Mazur et al., 2014; Mosaly & Mazur, 2011; Ramkumar et al., 2015).

Analytical evaluation methods are popular in HCI evaluation because they often require less formal training, take little time to perform, and can be used in both early and late stages of the development process. Models that quantify estimated workloads were often used in analytical evaluation. Previous research indicates that using models are more consistent and quantifiable than using individual measures. However, it should also be noted that accuracy of the model highly depends on the completion of the tasks and the time required for building such a model also depends on the complexity of the task.

For instance, the GOMS model is a specialized human information processor model for HCI observation. It is a method for describing a task and the user's knowledge of how to perform the task in terms of goals, operators, methods, and selection rules. Here *goals* refer to a particular state the user wants to achieve in their software or service. Goals are achieved by *methods*, which themselves contain operators that should be performed in a particular sequence to accomplish that goal. Methods are well-learned procedures for accomplishing the goals. A method consists of sequences of steps for accomplishing the goal. A classic example of "deleting a paragraph in a text editor" method can be described as: using a mouse, place the cursor at the beginning of the paragraph, push the mouse left button down, drag the cursor to the end of the paragraph, release the mouse left button, highlight the paragraph, then hit the delete key. Another (less efficient) method can be: place the cursor at the end of the paragraph and hit the delete key until the paragraph is gone. *Selection rules* are used to determine which method to select when there are more than one available at a given stage of a task.

Operators are the actions that are performed in using a method. With the original command-line interfaces, an operator was a command and its parameters, typed on a keyboard. In a graphical user interfaces, typical operators are menu selections, button presses, mouse clicks, etc. In some studies, gestures, spoken commands, or even eye movements are considered as operators (Lin, Hsieh, & Lin, 2013).

In 1983, Card et al. (Card, Moran, & Newell, 1983) initiated the study of GOMS by their CMN GOMS model. CMN GOMS has a strict goal hierarchy and methods are represented in an informal form and can include sub-methods. Apart from CMN GOMS, many other types of GOMS models have been discussed in the literature: the Keystroke-Level Model (KLM GOMS) (Kieras, 1993), the Natural GOMS Language (NGOMSL) model (Kieras, 1988, 1997a), the Cognitive Perceptual Motor (CPM) GOMS model (John & Kieras, 1996), and a more recent variation of GOMS named Sociotechnical GOMS (SGOMS) (West & Nagy, 2007). The KLM GOMS model is a simplified version of the CMN-GOMS model. It only utilizes six primitive operators as: 1) pressing a key; 2) moving the pointing device to a specific location; 3) pointer drag movements; 4) mental preparation; 5) moving hands to appropriate locations; and 6) waiting for the computer to execute a command. A more rigorously defined version of the KLM GOMS model is named the NGOMSL model (Kieras, 1988, 1997a) which presents a procedure for identifying all the GOMS components, expressed in a form similar to an ordinary computer programming language. The NGOMSL model includes rules-of-thumb about how many steps can be part of a method, how goals are set and achieved, and what types of information should be remembered by the user while doing the task. The CPM-GOMS model was introduced to describe parallel activities (John & Kieras, 1996). It utilizes cognitive, perceptual, and motor operators in a critical-path schedule chart to resemble multitasking behaviors of the user. West et al. (West & Nagy, 2007) developed Sociotechnical GOMS (SGOMS) model, which extends the idea of using a control structure for dealing with processes such as planning, scheduling, and teamwork from micro to macro level tasks. SGOMS consists of two components: the first part of SGOMS is the planning unit which is a sequence of unit tasks for accomplishing a specific goal, the second component of SGOMS is a framework that describes how planning units fit into the work process. Christou, Ritter, and Jacob (2012) developed a new GOMS model named *codein* to support the evaluation of reality based interaction styles. The main advantage of their GOMS model was that it was able to evaluate the task completion time of parallel actions during the performance of a task which was only possible using CPM-GOMS.

In the past decade, GOMS model has been extensively applied in developing analytic models of user behavior for user interaction evaluation. Carmel, Crawford, and Chen (1992) applied the GOMS model to analyze hypertext browsing strategies with a HyperCard application. They treated browsing as a cognitive information processing activity, and attempted to describe the browsing process both qualitatively and quantitatively. In their research, they identified three different types of browsing patterns: search-oriented, review and scan. In addition, they also compared tactics used by novice and expert users on a specific topic. Smelcer (1995) used a NGOMSL model to identify causes of user errors for database query composition. Saitwal, Feng, and Walji (2010) also used the GOMS model to evaluate the electronic health record (EHR) systems and proposed suggestions for improving user interfaces. GOMS has also been successfully used to determine the usability of websites for disabled users (Schrepp & Fischer, 2006), to measure the performance on how users interact with web applications (Andrés, 2014), to assess the performance of automobile human–machine interfaces (Xiang & Chen, 2010), and the navigational structure of websites (Oyewole & Haight, 2011). Although it was designed to predict task execution time on mouse and keyboard systems, the GOMS model is flexible enough to be adjusted to measure the HCI performance of using touch screens (Abdulin, 2011) as well.

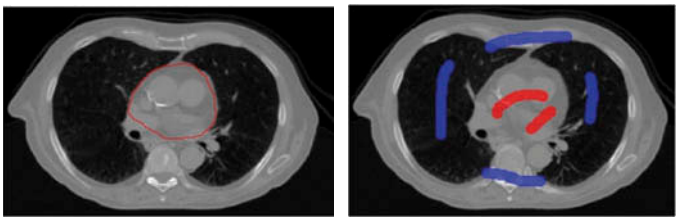
The literature survey indicates that the GOMS model and the NASA-TLX questionnaire have been used to identify the workload and performance of the users in many case studies. However, the

questions of what are the inter-relations between these two measures, and how to combine those measures to identify the design issues and offer design suggestions remain to be answered.

3. Prototype Design

For the proposed research, two interactive segmentation prototypes were developed as a plug-in on the Medical Imaging and Interaction Toolkit (MITK, 2016) platform. Though both prototypes share the same computational algorithm, two different types of HCI *inputs*, which are named as the *contour approach* and the strokes *approach*, were developed to help physicians provide their inputs to initialize the algorithm. (Dolz et al., 2014a). Using the contour approach, the user utilizes line based inputs and draws contours of an anatomical structure in a limited number of slices as shown in Figure 2a. The algorithm then computes contours of this structure in the rest slices. The contour approach is the most familiar method for users, as it is used in many types of clinical software (Aquilab Artiview®, 2016; Variance Medical Systems, Eclipse®, 2016). In this study, the mouse is utilized as the *input device* in the use of the contour approach. *Tools* which can be used for drawing and modifications can be selected from the panel at the right side of the window (Figure 3). In the contour approach, a free hand drawing tool is provided to the user. Besides, the user could also use a paintbrush tool where the brush size can be adjusted by a slide bar. In the interactive segmentation which utilizes the contour approach, physicians are instructed to draw the contours accurately on the slice they select. Hence this input can be physically and mentally demanding for the physician.

For the second prototype, an area based input approach, which is referred as *the strokes approach*, was introduced. This approach was designed to reduce the physical and mental demands of the users. The physician draws strokes to indicate the foreground (FG, as the two red strokes in Figure 2b) that represents the region the physician wants to include as an organ and the background (BG, as the four blue strokes in Figure 2b) that distinguishes the areas which should not be included in the organ. The algorithm then computes the segmentation volume. Using this approach, physicians may indicate the ROIs by short strokes or some dots. However, compared to the contour approach, inputs by the strokes approach are not widely used in interactive segmentation for radiotherapy planning. It is expected that using the strokes approach will result in: a) shorter drawing time, b) shorter thinking time, and c) introducing an extra swap of tools between FG and BG, which may lead to more HCI errors than using the contour approach. Using the strokes approach, the accuracy requirement of the interaction is not high, thus the paint brush is the only tool to facilitate the inputs.



(a) The *contour* approach (b) The *strokes* approach

Figure 2. Two designed interactive segmentation approaches.

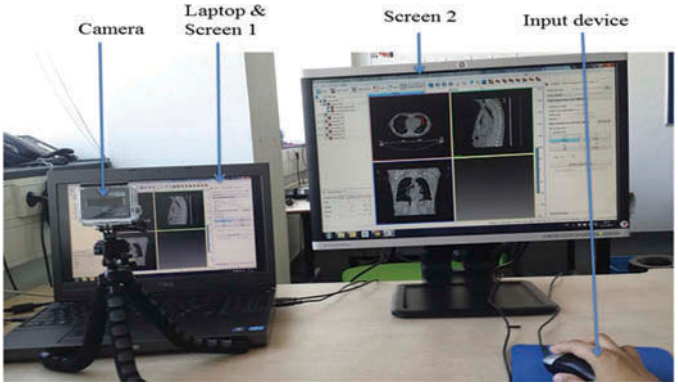


Figure 3. The user testing setup.

4. Experimental Setup

This study was conducted at the Department of Radiation Oncology, The University Medical Center Freiburg, Freiburg, Germany and Faculty of Industrial Design Engineering, Delft University of Technology, The Netherlands. Datasets of five patients who underwent planning CT (pCT) for lung cancer treatment were selected. Utilization of the datasets for this study was approved by the Ethics committee of The University Medical Center Freiburg, Freiburg, Germany. Three resident physicians joined the study. The physicians were asked to contour four different types of OAR, i.e., the spinal cord, the lungs, the heart, and the trachea using both prototypes, respectively. In the axial direction, the spinal cord and the trachea have a relatively small dimensions where the heart and lungs are larger (diameters of the spinal cord, trachea, heart and right lung in an axial plane are approximately 1–1.5 cm, 2.5 cm, 6.5–7 cm and 12–12.5 cm, respectively). Furthermore, the extents of those organs in the sagittal direction (the length) are different. For instance, the spinal cord is approximately 45 cm in length, while the heart is only 12 cm long. Hence the number of 2D CT image slices in the sagittal direction varies as well. Figure 3 shows the setup of the study where the prototypes were installed on a laptop. The laptop display (Screen 1) was mirrored on a 22-inch monitor (Screen 2), which is the screen size that physicians are familiar with. A camera was setup in front of the laptop screen to record the complete interaction process. The software also automatically logged some user interactions into a log file.

4.1. Analytical Measure of the Process

Based on video analysis, the use of each GOMS operators in HCI process and its duration were recorded. Apart from this we also measured the number of errors made during the whole segmentation process for each approach. Paired t-tests were used to identify if there are any statistically significant differences among the results.

4.2. Subjective Measure of the Process

In this experiment, the NASA-TLX questionnaire was used as a subjective measure to determine the workload of the user during the segmentation process. Physicians were asked to fill in the questionnaire each time after they finished a case. In addition, a

short interview was conducted to discuss the answers in the NASA-TLX questionnaire.

5. Result

5.1. GOMS Model

Goals

The top level goal of this task was to segment the organs at risk (OAR) using two types of user input approaches.

Operators

This study identified mainly 8 categories of operators: *mouse cursor move, zooming, panning, mouse click, scroll, draw* and *brush size adjustment*, that were used in segmenting the OAR. Among them, five categories only have one operator and the rest 3, *draw, scroll* and *mouse click*, can be further detailed. The *draw* category has *draw FG, draw BG* and *draw contour* operators; the *scroll* category also had three operators: *fast scroll, slow scroll* and *normal scroll*; the *click* category consists of five operators: *click FG, click BG, click paint, click add*, and *click wipe*. Among those operators, *draw FG, draw BG* and *click FG, click BG* are associated only with the strokes interaction. Table 1 shows the operators that were identified in this study for the two types of user inputs. The duration of each operator and the explanation of each operator are presented as well.

Methods

In many cases, a fixed combination of multiple operators was used in the HCI process to achieve a certain goal. For instance, *click paint* was followed by a *mouse move* and *draw* operators in order to segment a single slice. Those fixed combination of multiple operators are named methods. Ten different methods were identified in the use of both the input methods as shown in Figure 4. In the figure, the vertical axis indicates the operators that were used in the method while the horizontal axis indicates which step this operator was used in the method. At the right of the figure different types of method are explained. In all methods the first two interactions performed by physicians were usually *zooming* and *panning*. Hence, *zooming* and *panning* operators are not presented in the explanation. The next step which was observed in most of the methods was that physicians chose the tool and started contouring on the presented slice without scrolling to other slices, which indicates physicians' high confidence on the human anatomy. Only in three methods, physicians scrolled to different slices to provide their inputs. The scrolling time and the drawing time of each method may differ due to different numbers of slices scrolled and the dimensions of the organs, respectively.

Table 1. GOMS Operators.

No.	Operators	Time (s)	Meaning
1	*Mouse cursor move*	0.9	Moving of the cursor from the drawing region to a panel to select a tool
2	*Zooming*	2	Right mouse button down and move the mouse
3	*Panning*	2	Middle mouse button down and move the mouse
4	*Mouse clicks*	0.2	Left mouse button click
	Click paint		
	Click FG tool		
	Click BG tool		
	Click Add tool		
	Click wipe		
5	*Scrolling time*		Mouse wheel scroll forth and back
	Slow scroll	0.8	Observed during decision making process
	Normal scroll		
	Fast scroll	0.3	Observed when the user wanted to reach the target region
		0.03	Mainly observed while familiarization with the anatomy of the dataset
6	*Drawing time*		Left mouse button down. Drawing time differed between the organs, interaction methods and physicians. Hence there was no fixed time for drawing
	Draw FG		
	Draw BG		
	Draw contour		
7	*Wipe*	2–6	Left mouse button down Observed mainly when the user created mistakes
8	*Adjustment of the brush size*	0.4–2	Observed with the paint tool, mainly when the users shifted between tools

A workflow is a combination of different methods and operators to achieve a complete segmentation of an organ. The workflow can also be referred as a unit task, as unit tasks refer to the combination of a sequence of smaller tasks in order to achieve a global goal. Figure 5 shows two examples of workflows. From Figure 5, it can be observed that Workflow 1 is achieved using combinations of method 7 and 1 and Workflow 2 is achieved using method 4 and 1. Using different selection rules, the users may combine different methods together to form different workflows for achieving the same task.

5.2. Different Operators and Their Average Time during One Segmentation Process

Table 2 shows the total time taken by different operators during the whole segmentation process for both interactive segmentation

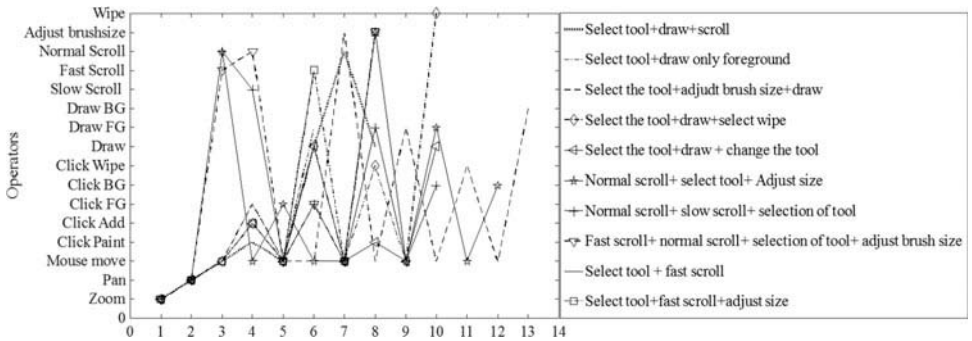


Figure 4. Ten different methods.

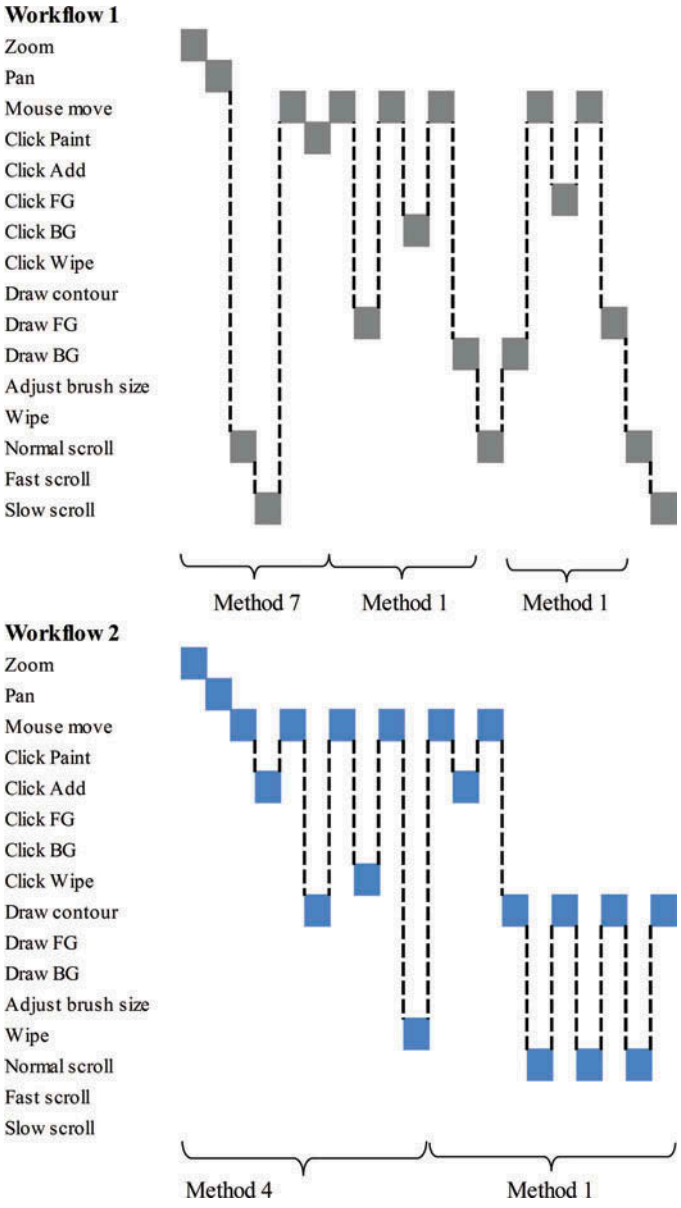


Figure 5. Examples of workflows (Workflow 1 is a combination of method 7 and 1, Workflow 2 is a combination of method 4 and 1).

approaches. As mentioned before, the dimensions of the organs are different and hence the overall time of using operators is different for each OAR. However, for each type of operator, except for the drawing time which is strongly associated to the dimensions, the average time taken is nearly same as Table 1. When the two approaches were compared against each other, for all the physicians, lung segmentation showed significant difference in the input time (p = 0.02) using a paired t-test, where the strokes approach was much faster than the contour approach. Even though there were differences in the mean segmentation time for other organs, these differences were not statistically significant.

5.3. NASA—TLX Questionnaire

Figures 6a and 6b show the individual workloads for the two types of approaches using NASA-TLX questionnaire. The overall workload is calculated by taking the average of all the individual workloads. The spinal cord and trachea shows higher workload for the contour approach, however the difference was not statistically significant. Only in lung segmentation, a statistically significant difference in the workload (p = 0.0002) between the two interactive segmentation approaches was identified.

5.4. Predicting NASA-TLX Using GOMS Operators

Using the linear regression method, we modeled the relations between the workloads identified using NASA-TLX questionnaires and the overall usage time durations of each GOMS operator. In the linear regression, the overall time durations of each of the six GOMS operators, i.e. *Draw, Slow scroll (SS), Normal Scroll (NS), Fast Scroll (FS), Mouse Move (MM)* and *Mouse Click (CLICK)* were used as predictors, and different types of workloads in the NASA-TLX questionnaire were used as criterion variables. Equations 1 and 2 show the models regarding the strokes approach and the contour approach, respectively. In the regression, the workloads of each physician measured by the NASA-TLX questionnaires were adjusted to a mean of 50 and the standard deviations for different types of workloads and for every physician were normalized as well.

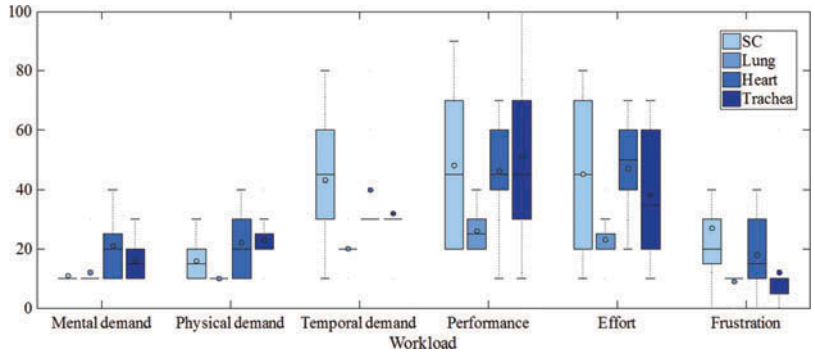
$$\begin{bmatrix} 0.5 & 1 & -0.1 & -0.7 & -0.1 & 2 & 6.9 \\ 0.5 & 0.2 & 0.06 & -1.1 & -0.08 & -5.3 & 18 \\ 0.7 & -0.1 & 1.1 & 0.1 & -0.1 & -7.2 & 18 \\ 0.6 & -0.6 & 0.6 & -0.7 & -0.1 & -8.7 & 50 \\ 0.9 & 0.1 & 0.7 & 1.9 & -0.7 & -8.3 & 20 \\ 0.4 & 0.6 & 0.4 & -0.4 & -0.06 & 4 & 1.3 \end{bmatrix} \begin{bmatrix} \text{DRAW} \\ \text{SS} \\ \text{NS} \\ \text{FS} \\ \text{MM} \\ \text{CLICK} \\ 1 \end{bmatrix} = \begin{bmatrix} \text{MENTALDEMAND} \\ \text{PHYSICALDEMANDTEMPORALDEMAND} \\ \text{PERFORMANCE} \\ \text{EFFORT} \\ \text{FRUSTRATION} \end{bmatrix} \quad (1)$$

$$\begin{bmatrix} 0.1 & -0.2 & -0.2 & -1.5 & -0.3 & -1 & 28 \\ 0.2 & -0.2 & -0.3 & 2.8 & -2.6 & 17 & 30 \\ 0.2 & -0.4 & 0.1 & -3.4 & 0.8 & -19.5 & 60 \\ 0.3 & -0.1 & -0.5 & -6.6 & 2.1 & -30.4 & 71.3 \\ 0.1 & -0.2 & 0.1 & -1.9 & 1.1 & -20.8 & 70 \\ 0.02 & 0.1 & 0.5 & -5 & 0.9 & -20.2 & 38 \end{bmatrix} \begin{bmatrix} \text{DRAW} \\ \text{SS} \\ \text{NS} \\ \text{FS} \\ \text{MM} \\ \text{CLICK} \\ 1 \end{bmatrix} = \begin{bmatrix} \text{MENTALDEMAND} \\ \text{PHYSICALDEMAND} \\ \text{TEMPORALDEMAND} \\ \text{PERFORMANCE} \\ \text{EFFORT} \\ \text{FRUSTRATION} \end{bmatrix} \quad (2)$$

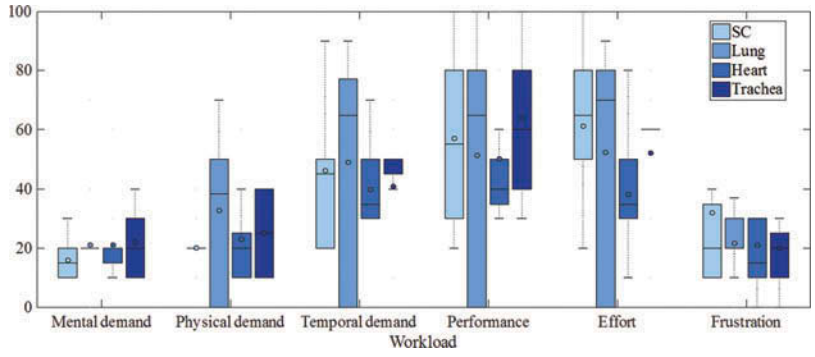
From the model, it can be identified that some predictors contribute significantly to one (or several) types of workloads (criterion variables) in the NASA-TLX questionnaire. For instance, the overall time durations of using the *draw* operator and the *slow scroll* (SS) operator are strongly associated to the mental demands when using either the strokes (significance level: 0.001 and 0.01) or the contour approaches (significance level: 0.03 and 0.02). The overall time durations of using the *draw, mouse click and mouse*

Table 2. The Average Time of GOMS Operators in Using Two Different Approaches.

Organs	Type of methods	Drawing time (sec)	Scrolling time (sec)	Normal scroll (sec)	Slow scroll (sec)	Fast scroll (sec)	Mouse moves (sec)	Click time (sec)
Spinal cord	Stroke	35.83 ± 14	36.13 ± 13	20.6 ± 14	7.2 ± 7	4.42	15.9 ± 7	1.3
	Contour	45.28 ± 19	27.33 ± 11.8	18.4 ± 11	2.9 ± 2	3.43 ± 2	13.3 ± 8	0.5
Lungs	Stroke	42.75 ± 7.9	19.24 ± 2	12.5 ± 1	4.2 ± 3	2.52 ± 1	24.3 ± 7	2.4
	Contour	219.75 ± 119	64.04 ± 57	62.5 ± 63	5.8 ± 3	3 ± 2	29.8 ± 20	3 ± 2
Heart	Stroke	65.7 ± 19	19.42 ± 14	15 ± 11	8.8 ± 5	0.4	14.2 ± 8	1.46 ± 1
	Contour	54.78 ± 18	25.64 ± 23	7.4 ± 7	17.4 ± 16	0.7	14.1 ± 10	1.4 ± 1
Trachea	Stroke	51.6 ± 14	13.78 ± 4.5	9.3 ± 3	2.8 ± 3	0	22.14 ± 12	1.68
	Contour	53.8 ± 14	16.86 ± 5.24	6 ± 2	10.8 ± 7	0	12.6 ± 7	1.4



(a) Using the strokes approach



(b) Using the contour approach

Figure 6. The outcomes of NASA-TLX questionnaires.

move operator are strongly associated with the physical demands (significance level: 0.01, 0.007, and 0.04). The time duration of using the *draw* operator and the *normal scroll* operator are also strongly associated with the temporal demand for strokes approach with significance levels of 0.004 and 0.01, respectively. For performance, effort and frustration, we did not find statistically significantly associated predictors.

5.5. Errors

Table 3 shows the common errors made by the physicians using the proposed two interactive segmentation approaches. A total of 58 errors were identified where 37 of them happened in using the strokes approach and the rest 21 belong to the contour approach. The most common error in using both the approaches was the wrong selection of tools, which contributes to 57% of the total errors. For instance, when the physicians chose the wipe tool they forgot to change it back to the paint tool. Instead they started giving the input using the same tool. The second most common error was in the selection of tools, with physicians sometimes clicking the same option twice resulting in deselection of the tools.

6. Discussion

In this section, we discuss the outcomes of the GOMS model and the NASA-TLX questionnaire in the evaluation of two interactive segmentation approaches for radiotherapy. First, we discuss the inter-relations between the GOMS model and the NASA-TLX questionnaire. Then the design suggestions regarding the two interactive segmentation approaches are proposed based on a synthesis of the outcomes of the GOMS model and the NASA-TLX questionnaire. Detailed suggestions, which mainly based on the outcomes of the GOMS model regarding each step of the HCI, are proposed as well.

6.1. Inter-Relations between GOMS Model and NASA-TLX

From Table 1 it can be seen that for both interaction approaches, we identified 8 main categories of GOMS operators where

Table 3. Percentage of Errors in Using Both Approaches.

Errors	Percentage of errors	
	Strokes	Contours
Paint and Wipe operator-With the Wipe tool the users drew on the image and with the Paint tool the users wiped the contour	33%	24%
Click operator-The tool was selected but the user clicked it again and deselected the tool by mistake	7%	10%
Zoom operator—Wrong zoom operations	12%	2%
Click FG and BG operator—Wrong selection of drawing tools (placed BG seeds instead of FG)	7%	–
Click operator—Users forgot to choose the paint tool option instead just selected the FG option	5%	–

drawing, scrolling and *mouse clicks* also have different variants. Besides, using NASA-TLX questionnaire we identified the workload of the users in using both approaches. In an earlier study conducted by Gao et al. (2015), there was not a single analytical measure that significantly correlated to the workloads in the NASA-TLX questionnaire. However, in this study we were able to identify some individual operators that contribute significantly to the workloads. According to Miyake (2001), an integrated objective measure is considered more reliable than using an individual measure. We also identified that using combination of measures predicted the workload better than using individual measures. For instance, it was better to predict the physical workload by combining the measures of *draw, NS, click* and *mouse cursor move operators* instead of just predicting using *draw* or *NS* operator only. The correlation coefficient between the *draw* operator and the physical demand is only 30%, however, by combining with other operators, the correlation coefficient rises to 60%.

Using regression analysis, we associated the GOMS operators to the mental, physical and temporal demands which were identified by the NASA-TLX questionnaires. Effort and performance demands could not be predicted well using either the individual or combined GOMS operators. A decrease in drawing time will decrease the workload of the users, which was confirmed by the low levels of physical and mental demand found with NASA-TLX using the strokes approach in lung segmentation. In our study, the performance measure on the NASA-TLX questionnaire include aspects of the HCI process while performing the task and are not just limited to the end result. Even after explaining this to users beforehand, the interviews after completion of the tasks indicated that the performance measure was heavily influenced by the end result instead of the HCI process, especially when the quality of the result differed. This partially explains that performance could not be predicted well using either the individual or combined GOMS operators. Hence, we recommend that in a result oriented task, the outcomes of the performance measure should be carefully analyzed.

To categorize different operators, we found that the *draw operator* is associated with both the physical and mental demands, hence it can be categorized as a semi-cognitive and semi-physical operator and the *mouse click* can be categorized as physical operator. The *slow scroll* operator contributed significantly to the mental demand in both scenarios. Based on this we concluded that *slow scroll* is more a cognitive operator than a physical operator. Unlike the mouse click operator, scroll operators identified in this study do not consist of a single task. Instead, it is a fairly complex unit task which may involve different motor, perceptual, and cognitive operators to build up the context. However, we did not have sufficient measures to clearly distinguish if it is a method or an operator. For instance, as we did not measure any eye-movements hence we could not derive which operator contributes to the perception operators in the CPM-GOMS model.

6.2. Design Issues

The Two Designed HCI Input Approaches

Based on the results of the GOMS model (Table 2), it can be seen that the designed strokes approach was faster in segmenting lungs. The average drawing and scrolling time by the strokes approach in lung segmentation is almost 75% less than the time taken by the contour approach. For the rest of the organs, there was no statistical significant difference in using both approaches. However, the strokes approach introduces an increased shifting between the FG and BG tools. Consequently, it led to 7% of the total errors.

These findings can be further confirmed by the results of the NASA-TLX questionnaires, especially regarding the associated demands. Except for lung segmentation, there was no statistically significant difference in the workload between the two approaches (Figure 6). It could be explained that the lung is the largest structure (diameters of the spinal cord, trachea, heart and lung in an axial plane were 2, 2.5, 6.5–7, and 12–12.5 cm, respectively). Hence, designing tools that are able to automatically identify the type of organ being segmented and adjust their properties accordingly are recommendations for future designs.

Other Design Issues

The GOMS model has the advantage that it can model the HCI process in a continuous manner where the NASA-TLX questionnaire can only identify the workload of the HCI process at the end of the study (Bruneau, 2006). Thus, from GOMS we were able to identify more detailed design issues than from the NASA-TLX questionnaire. From Table 1, it can be identified that the time taken for operators such as *click* and *release mouse button* is in accordance with the literature (Kieras, 1993). The operator *Mouse cursor move* took on average 0.2 seconds, which was less than reported in literature (Kieras, 1993). This may be explained by differences in the mouse travel distance in the graphical user interface.

Table 3 shows that switching between the wipe and the drawing tool contributes to 57% of the errors. This was mainly seen in method 4. The wipe tool was used when a mistake was made or physicians were not satisfied with what they drew. One way to solve this issue could be that integrating opposite functionalities in one tool, e.g. using a "Nudge" tool, where the user can enlarge the contour by pushing contour from inside and using the same tool, the use can shrink the contour by pushing it from outside. This will help to reduce the frequency of changing tools. As a result, the distance of mouse movement and the numbers of mouse clicks will drop, which will save time and also reduce the number of errors.

Three different scroll (slow, normal and fast) operators were identified using GOMS model and it was mainly

observed in method 6, 7, and 8. In these three methods, the user scrolled through the dataset either to the start or to the end of the dataset. The *slow scroll* operator was mainly observed when the physicians were making decisions to choose the right slice to provide their inputs by comparing the anatomy and the contour they drew in the previous slice. Hence the time required for this method was longer than others and it involved a lot of decision making processes. This method was observed mainly in segmenting the heart and the trachea. In the case of heart segmentation, at the start of the procedure the physicians do not have the context from the previous or the next slice, so they have to scroll forth and back in order to check the contours and to take the right decisions about the anatomy/structure. In the case of trachea segmentation, the physicians compared the contours to the previously drawn contour in order to include the cartilage. The design suggestions are that the system can propose a contour on the current slice by considering the previously drawn contours, or two small windows can be designed to show the previous contoured slice and the next slice to be contoured.

7. Limitations

One of the limitations of this study is that only three experts participated in the study. For a specialized domain such as radiation oncology, it is difficult to organize a large number of experts as the required expertise is very specific and a considerable amount of time is required for each physician during the pilot, the main experiments and the interviews, etc. Thus, the outcomes from this study are more suggestions for improvement. Besides for some operators, a more in-depth analysis is needed for a more detailed GOMS model with the help of more measures. This will be considered in our future work.

8. Conclusions

In this study, we used the GOMS model and the NASA-TLX questionnaire to evaluate the HCI process and to propose design suggestions for interactive segmentation in radiotherapy. Using the GOMS model we identified 16 different operators and ten different methods that were involved in the segmentation process. Those operators can be further associated to the mental, physical and temporal demands, identified by NASA-TLX questionnaire using regression analysis. The significance of predictors in the regression analysis also helped us identify that if a GOMS operator was a cognitive or physical operator according to its associated demands in the NASA-TLX.

Regarding the segmentation process, the designed strokes approach was faster and less demanding in segmenting large organs based on the findings and inter-relations between the GOMS operators and the results of the NASA-TLX questionnaire. However, it introduces an increased number of shifts between different HCI tools. As a result, physicians tended to make more errors than using the traditional contour approach. For smaller organs, there was no statistical significant difference in using both approaches. Hence, designing tools that automatically identify the organ being segmented and adjust their properties accordingly are recommendations for future designs. Besides, new HCI tools which are able to integrating opposite functions, should be considered as well.

Future study should also focus on involving more HCI components, e.g., new input devices and tools, in order to identify their effects on the HCI process and the segmentation results. More physicians will be involved in the experiment. In addition, more types of subjective, physiological and analytical measures will be incorporated in order to identify the relations among those measurements for offering better design suggestions.

Acknowledgments

The authors would like to thank Mr. Jose Dolz and Dr. Hortense A Kirisli from Aquilab, France, for helping in the implementation of the prototypes. The authors would also like to express their appreciations to other members of the SUMMER consortium for their valuable advices regarding the proposed research.

Funding

The presented research is part of Software for the Use of Multi-Modality images in External Radiotherapy (SUMMER) project which is funded by European Commission (FP7-PEOPLE-2011-ITN) under grant agreement PITN-GA-2011-290148.

References

Abdulin, E. (2011). *Using the keystroke-level model for designing user interface on middle-sized touch screens*. Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems (pp. 673–686). Vancouver, Canada.

Andrés, J. D. (2014). Towards an automatic user profiling system for online information sites Identifying demographic determining factors. *Online Information Review*, *39*(1), 61–80.

Aquilab Artiview. (2016). A complete software platform for multimodality imaging, contouring, and evaluation in radiotherapy. Retrieved April 1, 2016, from http://www.aquilab.com/index.php/artiview.html

Aselmaa, A., Goossens, R. H., Laprie, A., Ken, S., Fechter, T., Ramkumar, A., & Freudenthal, A. (2013). *Workflow Analysis Report*. Retrieved March 31, 2016, from http://summerproject.eu/work/deliverables/

Balafar, M., Ramli, A., Saripan, M., & Mashohor, S. (2010). Review of brain MRI image segmentation methods. *Artificial Intelligence Review*, *33*(3), 261–274.

Boykov, Y., & Jolly, M. P. (2001). *Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images*. International Conference on Computer Vision (pp. 105–112). Vancouver, Canada.

Bridger, R. S., & Brasher, K. (2011). Cognitive task demands, self-control demands and the mental well-being of office workers. *Ergonomics*, *54*(9), 830–839.

Bruneau, D. P. J. (2006). *Subjective mental workload assessment, International encyclopaedia of ergonomics and human factors* (2nd ed., pp. 946–947). 1, USA: CRC press.

Cain, B. (2007). A *review of the mental workload literature* (Report #RTOTR-HFM-121-part-II). Defense Research and Development, Toronto, Canada. Retrieved March 15, 2016, from http://www.dtic.mil/dtic/tr/fulltext/u2/a474193.pdf.

Card, S., Moran, T., & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates. ISBN:0898592437.

Carmel, E., Crawford, S., & Chen, H. (1992). Browsing in hypertext: A cognitive study. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(5), 865–884.

Chang, W., Hwang, W., & Ji, Y. G. (2011). Haptic seat interfaces for driver information and warning systems. *International Journal of Human-Computer Interaction*, *27*(12), 1119–1132.

Christou, G., Ritter, F. E., & Jacob, R. J. K. (2012). Codein-A new notation for GOMS to handle evaluations of reality-based interaction style interfaces. *International Journal of Human-Computer Interaction*, *28*(3), 189–201.

De Boer, R., Vrooman, H., Ikram, M., Vernooij, M., Breteler, M., Van de Lugt, A., & Niessen, W. (2010). Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods. *NeuroImage*, *51*(3), 1047–1056.

Dey, A., & Mann, D. D. (2010). Sensitivity and diagnosticity of NASA-TLX and simplified SWAT to assess the mental workload associated. *Journal of Ergonomics*, *53*(7), 848–857.

Dirican, A. C., & Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in Human Computer Interaction. *Procedia Computer Science*, *3*, 1361–1367.

Dolz, J., Kirisli, H. A., Viard, R., & Massoptier, L. (2014a). *Combining watershed and graph cuts methods to segment organs at risk in radiotherapy*. Proc. SPIE 9034, Medical Imaging: Image Processing. San Diego, CA.

Dolz, J., Kirisli, H. A., Viard, R., & Massoptier, L. (2014b). *Interactive approach to segment organs at risk in radiotherapy treatment planning*. Proc. SPIE 9034, Medical Imaging: Image Processing.

Gao, Q., Wang, Y., Song, F., Li, Z., & Dong, X. (2015). Mental workload measurement for emergency operating procedures in digital nuclear power plants. *Ergonomics*, *56*(7), 1070–1085.

Hammers, A., Chen, C. H., Lemieux, L., Allom, R., Vossos, S., Free, S. L., … Koepp, M. J. (2007). Statistical neuroanatomy of the human inferior frontal gyrus and probabilistic atlas in a standard stereotaxic space. *Human Brain Mapping*, *28*, 34–48.

Harders, M., & Székely, G. (2003). Enhancing human – Computer interaction in medical segmentation. *Proceedings of the IEEE*, *91*(9), 1430–1442.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human mental workload*. Amsterdam, Netherlands: North-Holland Press.

Heckel, F., Moltz, J. H., Tietjen, C., & Hahn, H. K. (2013). Sketch-based editing tools for tumour segmentation in 3D medical images. *Computer Graphics Forum*, *32*(8), 144–157.

Jeon, M., & Croschere, J. (2015). *Sorry, I'm late; I'm not in the mood: Negative emotions lengthen driving time*. 12th International Conference on Engineering Psychology and Cognitive Ergonomics (Vol. 9174, pp. 237–244). Los Angeles, California.

John, B. E., & Kieras, D. E. (1996). Using GOMS for user interface design and evaluation: Which technique? *ACM Transactions on Computer-Human Interaction*, *3*(4), 287–319.

Kang, Y., Engelke, K., & Kalender, W. A. (2004). Interactive 3D editing tools for image segmentation. *Medical Image Analysis*, *8*, 35–46.

Karasev, P., Kolesov, I., Fritscher, K., Vela, P., Mitchell, P., & Tannenbaum, A. (2013). Interactive medical image segmentation using PDE control of active contours. *IEEE Transactions on Medical Imaging*, *32*(11), 2127–2139.

Kieras, D. E. (1988). Towards a practical GOMS model methodology for user interface design. In M. Helander (Ed.), *The handbook of human-computer interaction* (pp. 135–158). North-Holland, Amsterdam: Elsevier.

Kieras, D. E. (1993). *Using the keystroke-level model to estimate execution times*. Unpublished Report. University of Michigan. Retrieved March 31, 2016, from http://www.pitt.edu/~cmlewis/KSM.pdf

Kieras, D. E. (1997a). A Guide to GOMS model usability evaluation using NGOMSL. In M. Helander, T. Landauer, & P. Prabhu (Eds.), *The handbook of human-computer interaction* (2nd ed., pp. 733–766). North-Holland, Amsterdam.

Kotani, K., & Horii, K. (2003). An analysis of muscular load and performance in using a pen-tablet system. *Journal of Physiological Anthropology*, *22*(2), 89–95.

Li, J. Y., & Dang, J. W. (2012). Research and improvement of live-wire interactive algorithm for medical image segmentation. *Applied Mechanics and Materials*, *182–183*, 1065–1068.

Lin, C. J., Hsieh, T. L., & Lin, S. F. (2013). Development of staffing evaluation principle for advanced main control room and the effect on situation awareness and mental workload. *Nuclear Engineering and Design*, *265*, 137–144.

Longo, L., & Kane, B. (2011, June 27–30). *A novel methodology for evaluating user interfaces in health care*. 24th International Symposium on Computer-Based Medical Systems(CBMS) (pp. 1–6). Briston, United Kingdom.

Lundström, C., Rydell, T., Forsell, C., Persson, A., & Ynnerman, A. (2011). Multi-touch table system for medical visualization: Application to orthopedic surgery planning. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 1775–1784.

Mazur, L. M., Mosaly, P. R., Hoyle, L. M., Jones, E. L., Chera, B. S., & Marks, L. B. (2014). Relating physician's workload with errors during radiation therapy planning. *Practical Radiation Oncology*, *4*(2), 71–75.

McGuinness, K., & O'Connor, N. E. (2010). A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, *43*(2), 434–444.

McGuinness, K., & O'Connor, N. E. (2011). Toward automated evaluation of interactive segmentation. *Computer Vision and Image Understanding*, *115*(6), 868–884.

Miyake, S. (2001). Multivariate workload evaluation combining physiological and subjective measures. *International Journal of Psychophysiology*, *40*(3), 233–238.

Morgan, J. F., & Hancock, P. A. (2011). The effect of prior task loading on mental workload: An example of hysteresis in driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(1), 75–86.

Mortensen, E. N., & Barrett, W. A. (1998). Interactive segmentation with intelligent scissors. *Graphical Models and Image Processing*, *60*(5), 349–384.

Mosaly, P. R., & Mazur, L. M. (2011). Empirical evaluation of workload of the radiation oncology physicist during radiation treatment planning and delivery. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 753–775.

Murata, A. (2006). Eye-gaze input versus mouse: Cursor control as a function of age. *International Journal of Human-Computer Interaction*, *21*(1), 1–14.

Olabarriaga, S., & Smeulders, A. (2001). Interaction in the segmentation of medical images: A survey. *Medical Image Analysis*, *5*(2), 127–142.

Oyewole, S. A., & Haight, J. M. (2011). Determination of optimal paths to task goals using expert system based on GOMS model. *Computers in Human Behavior*, *27*, 823–833.

Petitjean, C., & Dacher, J. (2011). A review of segmentation methods in short axis cardiac MR images. *Medical Image Analysis*, *15*, 169–184.

Ramkumar, A., Dolz, J., Kirisli, H. A., Varga, E., Schimek-Jasch, T., Nestle, U., … Song, Y. (2015). User interaction in semi-automatic segmentation of organs at Risk: A case study in radiotherapy. *Journal of Digital Imaging*, *29*(2), 264–277.

Roscoe, A. H., & Ellis, G. A. (1990). *A subjective rating scale for assessing pilot workload in flight: A decade of practical use* (Technical report TR 90019). Retrieved April 1, 2016, from

Rubio, S., Diaz, E., Martin, J., & Puente, J. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology*, *53*(1), 61–86.

Saitwal, H., Feng, X., & Walji, M. (2010). Assessing performance of an electronic health record (EHR) using cognitive task analysis. *International Journal of Medical Informatics*, *79*(7), 501–506.

Schrepp, M., & Fischer, P. A. (2006). GOMS model for keyboard navigation in web pages and web applications. *Lecture Notes in Computer Science*, *4061*, 287–294.

Sherbondy, A. J., Holmlund, D., Rubin, G. D., Schraedley, P. K., Winograd, T., & Napel, S. (2005). Alternative input devices for efficient navigation of large CT angiography data sets. *Radiology*, *234*, 391–398.

Smelcer, J. B. (1995). User errors in database query composition. *International Journal of Human–Computer Studies*, *42*, 353–381.

The Medical Imaging Interaction Toolkit (MITK). (2016). A toolkit facilitating the creation of interactive software by extending VTK and ITK. Retrieved April 1, 2016, from http://www.mitk.org

Varian Medical Systems, Eclipse. (2016). Treatment Planning System. Retrieved April 1, 2016, from https://www.varian.com/oncology/products/software/treatment-planning/eclipse-proton

Von Falck, C., Meier, S., Jördens, S., King, B., Galanski, M., & Shin, H. O. (2010). Semi-automated segmentation of pleural effusions in MDCT datasets. *Academic Radiology*, *17*(7), 841–848.

West, R. L., & Nagy, G. (2007). Using GOMS for modeling routine tasks within complex sociotechnical systems: Connecting macrocognitive models to microcognition. *Journal of Cognitive Engineering and Decision Making*, *1*(2), 186–211.

Xiang, L. X. L., & Chen, X. L. (2010). *The research on performance of automobile human-machine interface based on BHR-GOMS behavior model*. Intelligent computer intelligent systems (ICIS), IEEE international Conference (Vol. 2, pp. 174–178). Xiamen, China.

Yang, W., Cai, J., Zheng, J., & Luo, J. (2010). User-friendly interactive image segmentation through unified combinatorial user inputs. *IEEE Transactions on Image Processing*, *19*(9), 2470–2479.

Zhao, F., & Xie, X. (2013). An overview of interactive medical image segmentation. *Annals of BMVA*, *2013*(7), 1–22.

About the Authors

Anjana Ramkumar is a PhD researcher at the Department of Design Engineering, Faculty of Industrial Design Engineering, Delft University of Technology. Her main research interests are Human–computer interaction and user experience design.

Pieter Jan Stappers is professor of Design Techniques, focusing on tools and techniques to support designers in the early phases of the design process. He has published extensively on the topics of user research, especially on 'contextmapping', and research through design methodology

Wiro J. Niessen is a professor at Department of Medical Informatics at Erasmus MC, and Faculty of Applied Sciences of Delft University of Technology. He is a founding member of the Dutch Young Academy and is leading the Biomedical Image Analysis Platform of the European Institute of Biomedical Imaging Research.

Sonja Adebahr is a resident at the Department of Radiation Oncology at The University Medical Center Freiburg since 2005. For the past 4 years she has been working as research physician focusing on stereotactic body and conventional radiotherapy of lung tumors and 4D-Imaging.

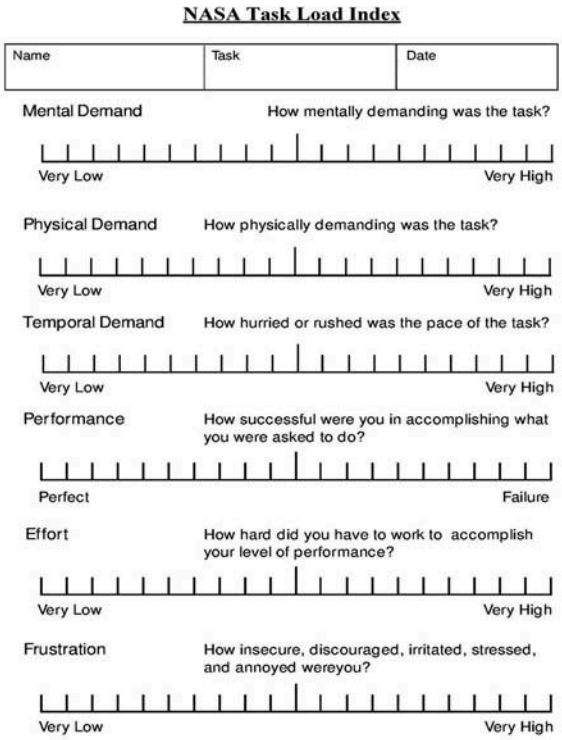
Tanja Schimek-Jasch is a resident in the Department of Radiation Oncology at The University Medical Center Freiburg since 2005. Her main research interests are the conception and implementation of clinical trials involving radiation oncology.

Ursula Nestle is an experienced radiologist and deputy director of Department of Radiation Oncology at The University Medical Center Freiburg. She specializes in radiation oncology and nuclear medicine. On top of her clinical obligations, Prof. Dr. Nestle was awarded an adjunct professorship at Albert-Ludwigs University Freiburg in 2012.

Yu Song is an assistant professor at the Department of Design Engineering, Faculty of Industrial Design Engineering, Delft University of Technology. His main research interests are 3D/4D image acquisition, reasoning, manipulation and Human–Computer Interaction.

Appendix A

An example of the NASA-TLX questionnaire (courtesy of Hart & Staveland, 1988)



Mental Demand : How much mental activity was required (deciding, thinking, calculating, remembering, looking, searching etc.)?

Physicial Demand : How much physicial activity was required (pushing, pulling, turning, controlling, activating etc.)?

Temporal Demand : How much time pressure did you feel during the task?

Performance : How successful you think you were in completing the goals?

Effort : How hard did you have to work (physically and mentally)?

Frustration : How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed, complacent did you feel during the task?