Customizable Gesturing Interface for the Operating Room using Kinect

Ali Bigdelou Tobias Benz Loren Schwarz Nassir Navab
Technische Universität München, Germany
ali.bigdelou@cs.tum.edu

Abstract

Sterility requirements and interventional workflow in the Operating Room (OR) often make it challenging for surgeons to interact with computerised systems. As a solution, we propose a customizable gesture-based interface for the OR. Our solution consists of a gesture recognition technique and an integration platform. Our recognition technique simultaneously detects the type of the gesture (categorical information) and the state that a user holds within the current gesture (spatio-temporal information). Introducing several feature extraction methods this technique performs independent to the human body proportions and Kinect placement. To exploit this gesturing technique in the OR, we additionally introduce an extensible software platform which allows to define context-aware gesturing interface for several intra-operative devices. The behavior of the gesturing interface can be customized using a visual editor.

1. Introduction

The operating room (OR) is a highly complex environment and surgical staff often work under high pressure. For assistance, various types of computerized systems are used. However, sterility regulations restrict the use of classical mouse-driven or touch-based interfaces and control terminals are often spatially separated from the main operating site [3]. Moreover, the surgeon might only have very limited freedom of movement while handling medical instrumentation. Typical solutions include delegating physical control over computerized systems to less-skilled assistants. This increases the amount of verbal communication, raising the chance of misunderstandings. The added level of indirection can adversely affect precision and efficiency [3]. As a solution we propose a gesturing interface for the OR [2]. Using this technique gestures can be customized according to the requirements of a specific surgery and surgeons do not have to adhere to a pre-defined set of movements that might interfere with their natural behavior in the OR.

2. Gesture Recognition Method

We propose a practical technique for gesture recognition and tracking based on skeletal data obtained from the Kinect body tracker. The method consists of an offline training step, where gesture models are learned from features extracted from the original data, and an online step, where the models are used to recognize gesture type and gesture state from previously unseen human poses. This method is simultaneously categorical and spatio-temporal [2].

2.1. Feature Extraction

In general, the performance of a recognition system primarily depends on defining suitable features to represent the input data. We propose a feature extraction method which consists of a feature representation and a feature normalization parts. We introduced three different feature representation techniques as *Distances Representation*, *Displacements Representation* and *Hierarchical Representation* [2]. These make the gesture-based interface independent from the users location and orientation as well as the placement of the Kinect sensor itself in the scene. We additionally proposed two schemes for normalization of these extracted features for independence of body styles and proportions as *Relative* and *Unit* normalizations [2]. These can improve the success rate of the recognition approach especially when the trainer user, is not the end-user of the system.

2.2. Learning and Recognizing Gestures

In the training phase, our method learns prior models of gestures from sample pose feature data. Let $\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ be a dataset of n feature vectors for N gestures of interest. Each vector $\mathbf{s}_i \in \mathbb{R}^d$ represents a particular body pose and is computed using either of the feature extraction and normalization methods discussed above. The known gesture labels for the training data are denoted by $\mathbf{C} = \{c_1, \ldots, c_n\}$, with $1 \leq c_i \leq N$. Using PCA for dimensionality reduction, we construct a one-dimensional intermediate representation $\mathbf{X} = \{x_1, \ldots, x_n\}$ from the training features, such that every $x_i \in \mathbb{R}$ corresponds to one $\mathbf{s}_i$. The resulting dimensionality-reduced datasets $\mathbf{X}^c$

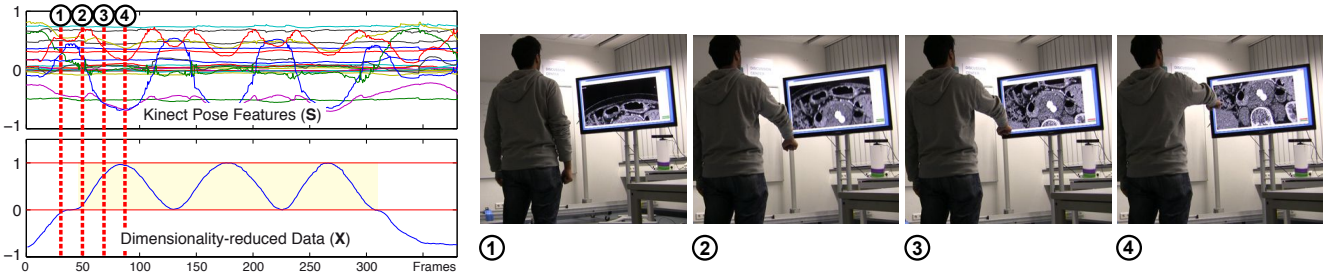

Figure 1. Exemplary scrolling gesture (*right*), highlighted in a stream of features (*left, top*) and one dimensional data (*left, bottom*).

are then normalized such that, for each gesture, the *control poses* map to the range $[0, 1]$ (see Figure 1).

After training, we are given new, previously unseen feature vectors $\mathbf{s}_t$ at any time t. Our aim is to determine the intermediate representation $\hat{x}_t \in \mathbb{R}$ corresponding to $\mathbf{s}_t$ and to identify the performed gesture $\hat{c}_t$, $1 \leq \hat{c}_t \leq N$. For this purpose, we define a kernel regression mapping that projects arbitrary feature vectors to the dimensionality-reduced representation $\mathbf{X}$. We predict the value $\hat{x}_t$ for a feature $\mathbf{s}_t$ as

$$\hat{x}_t = f(\mathbf{s}_t) = \sum_{i=1}^{n} \frac{w_i(\mathbf{s}_t)}{\sum_{j=1}^{n} w_j(\mathbf{s}_t)} \cdot x_i. \qquad (1)$$

The weights are given by a Gaussian kernel $w_i(\mathbf{s}_t) = \mathrm{k}(\mathbf{s}_t, \mathbf{s}_i) = \exp(-\frac{1}{2}\|(\mathbf{s}_t - \mathbf{s}_i)/\sigma\|^2)$ with a width σ derived from the variance of the training features $\mathbf{S}$. Using the weights, we also determine the current gesture index (type) $\hat{c}_t$ as $\hat{c}_t = c_{k_t}$, where$k_t = \arg\max_i w_i(\mathbf{s}_t)$.

3. Application in Practice

In order to exploit this gesturing technique in the OR we introduce a modular and extensible platform based on component-based architecture. Four main types of components are defined as *Input*, *Target*, *Gesture Recognition* and *Demultiplexer* components [2]. Additionally, utilizing a data pipeline design pattern, the extracted categorical and spatio-temporal data can be connected to the features of the target devices. Such interaction pipeline defines the behavior of the interface. This definition can be provided separately for each stage of the surgical workflow, making the interface context-aware. For example, same gesture can be used for different purposes during surgery. The interaction pipeline can be presented as a graph. Using such a representation, we have developed a visual editor for customization of the interaction pipeline, Figure 2.a.

4. Experiments and Conclusion

We have performed a detailed quantitative evaluation using all feature extraction methods [2]. A general observation is that maximum recognition rates decrease with an increasing total number of gestures. While this behavior is

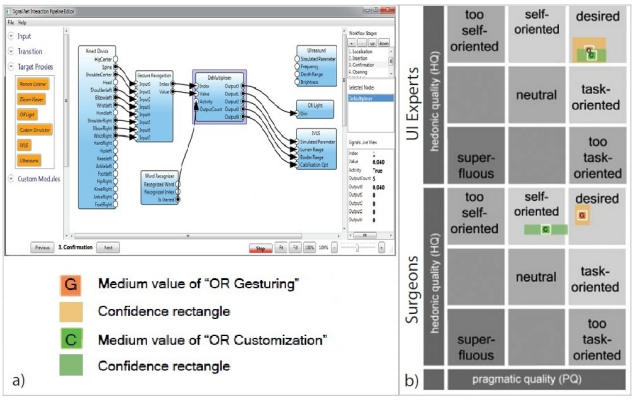

Figure 2. a) Visual pipeline editor, b) AttrakDiff HQ-PQ diagrams.

expected, the rates only drop slightly from above 90% in the case of 4 gestures to a reasonable 85% for the case of 16 gestures. In terms of feature selection, the best performance is achieved by the displacement features with unit normalization. We additionally conducted user studies with two groups of potential users, namely surgeons and user interface (UI) experts. Generate AttrakDiff HQ-PQ diagrams [1] are shown in Figure 2.b. All the surgeons agreed about the novelty, applicability and attractiveness of the gesturing interface. We noticed that learning to operate the visual editor is challenging for surgeons. UI experts have the impression that the customization concept is practical and straightforward. Based on them the system is fun to use and the learnability and rememberability of both interfaces are rated more than 80 percent. Overall, our experiments demonstrated the need for such approaches from surgeons perspective and proved the applicability of the proposed customizable gesturing method.

References

[1] AttrakDiff. http://www.attrakdiff.de/en/Home/.

[2] A. Bigdelou, T. Benz, L. Schwarz, and N. Navab. Simultaneous categorical and spatio-temporal 3d gestures using kinect. In *IEEE Symposium on 3D User Interfaces (3DUI)*, 2012.

[3] R. Johnson, K. O'Hara, A. Sellen, C. Cousins, and A. Criminisi. Exploring the potential for touchless interaction in image-guided interventional radiology. *Human Factors*, 2011.