Multimed Tools Appl
DOI 10.1007/s11042-014-1933-2

The impact of hesitation, a social signal, on a user's quality of experience in multimedia content retrieval

Tomaž Vodlan · Marko Tkalčič · Andrej Košir

Abstract The social signal (SS) of hesitation is commonly manifested through a multiplicity of nonverbal behavioural cues when a user is faced with a variety of decision choices. The aim of this study is to show that the utilization of the SS of hesitation in a conversational recommender system (RS) can improve the user quality of experience (QoE) when interacting with a video-on-demand system. An appropriate experimental design was modelled to detect the impact of the SS. The experimental scenario was a manual video-on-demand system with a conversational RS where the user selected one video clip among several presented on the screen. The system adjusted the list of the video items to be recommended according to the extracted SS class {hesitation, no hesitation}. To detect if the user was hesitating, we used hand movements, eye behaviour and time between two selections. Two user groups were tested to allow realistic estimation of the impact of the SS. In the user test group, the SS of hesitation was considered, while in the control group it was not. The evaluation of impact of the SS on QoE was based on pre- and post-interaction questionnaires. Our results showed a significant difference in user satisfaction with the system between those two groups, indicating that the use of SS of hesitation in conversational RS improves the QoE when the user interacts with a video-on-demand system.

Keywords Social signals · Hesitation · Human–computer interaction · Video-on-demand · Recommender system

1 Introduction

Social signals (SSs) have received much attention in recent years because they provide additional natural information about human behaviour, which offers important benefits in

T. Vodlan (✉)
Agila d.o.o., Tehnološki Park 19, 1000 Ljubljana, Slovenia
e-mail: tomaz.vodlan@gmail.com

M. Tkalčič
Johannes Kepler University, Linz, Austria
e-mail: marko.tkalcic@jku.at

A. Košir
Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
e-mail: andrej.kosir@ldos.fe.uni-lj.si

Published online: 19 March 2014

human–computer interaction (HCI) [44]. SSs are manifested through a multiplicity of nonverbal behavioural cues (e.g., gestures, postures, facial expressions), and present human reactions to current social situations. However, it is not clear how to utilize SSs in HCI applications, which is the major reason why the most systems in the HCI domain are socially ignorant. There are only few examples of using SSs in HCI. For example, Ferreira et al. [12] used user's SSs during interaction with a photocopier to predict the task difficulty. However, although the results of different studies [54–56] have shown that SS can provide important additional information regarding interaction between two people, little attention has been paid to utilization of SSs in HCI.

In this paper, we present the experimental design for measuring the impact of a particular SS, hesitation, on a user's decision when using a video-on-demand (VoD) system. The SS of hesitation is manifested frequently through behavioural cues when a user interacts with video items. The information about the produced SS is exploited by a recommender system (RS) to suggest relevant new videos to the end user. The impact of SS on the user decisions is measured based on the evaluation of pre- and post-interaction questionnaires of two user groups: the test group (SS is considered) and the control group (SS is not considered). Fisher's exact test, Mann–Whitney U test and independent t-test are used to compare the control and test group results from the questionnaires. The results demonstrate that the utilization of SS in the RS yielded a significantly higher quality of experience (QoE), which reflects user satisfaction with the system. During user interaction with the VoD system, the video selection time was also measured. The results showed that the utilization of SS does not reduce the video selection time.

The goals of this paper are (i) to introduce an experimental design for the evaluation of the impact of SS in a VoD system, (ii) to show that the use of the SS of hesitation in the RS can improve user's QoE when the user is interacting with a VoD system, and (iii) to show that the utilization of SS in RS does not reduce video selection time.

1.1 Motivation

The proposed methodology of our study is based on the use of the SS of hesitation in a VoD system as illustrated in Fig. 1. We address the usage scenario of the recommendation of relevant video content in a VoD system. State-of-the-art RSs present a solution, but they ignore the social context of the user. As such, why don't we use the SSs that the user produces when he interacts with the system as input information for a personalized RS? It has been shown that the accuracy of the quality of recommendations increases when implicit signals from the user (affect) are included in the RS [52]. Humans naturally produce SSs in several verbal and nonverbal ways [55]. Based on this, we can assume that user feedback SS can improve the user experience and increase the user satisfaction with a communication service. This is the main reason to research the impact of these naturally user-produced behavioural cues in HCI.

1.2 Why was the SS of hesitation selected?

As mentioned previously, we measured the impact of the SS of hesitation on user decisions. When we observe a user's interaction with the VoD system, several different behavioural cues could be observed, representing different SSs, with hesitation being one of them. The main reason why we use it in our design is that the assumption that the SS of hesitation is strongly correlated with the satisfaction of a user with the presented items [11]. This is then used as implicit feedback in a conversational RS (see Fig. 1). Conversational RS should narrow the set

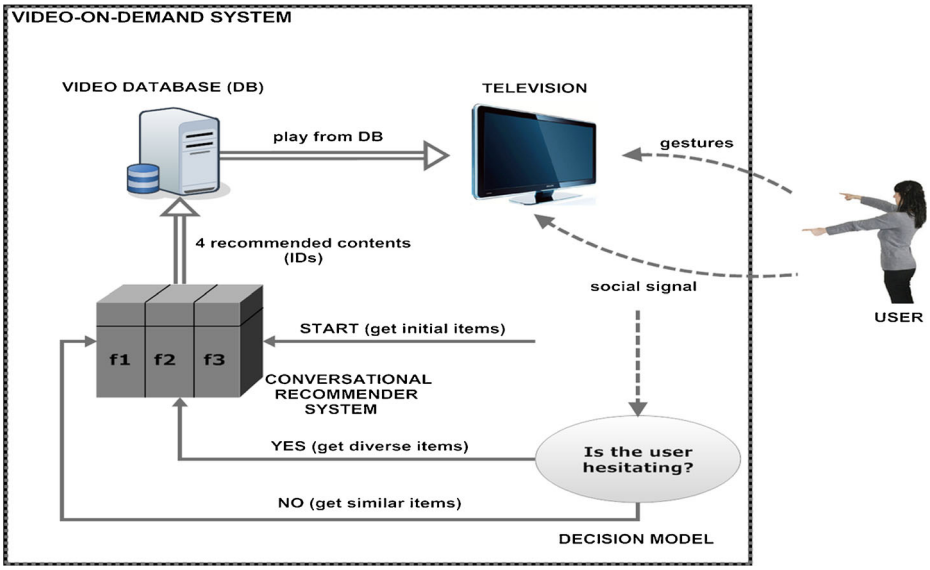

Fig. 1 The proposed design of our solution. SS of hesitation is used as input information for conversational RS. If the user hesitates, the conversational RS provides diverse new items (function f2). If the user is not hesitating, it provides similar new items (function f1). Function f3 provides video contents for the first screen (the first four videos that are projected on the screen when a user turns on the system). The video database contains the movie trailers that are played on screen

of returned items to the most relevant ones. Similar items should be selected when the user is not hesitating and diverse items should be selected when the user is hesitating (see sub-Subsection 5.3.5 for details).

1.3 Contribution of the paper

The main contribution of this paper is a proposed experimental design for the evaluation of the impact of SS of hesitation on the user's decisions during their interaction with a VoD system. Other contributions are: (i) a study of behavioural cues that are most significant for the SS of hesitation and (ii) a study of impacts on the QoE when the user interacts with a VoD system.

1.4 Organization of the paper

The remainder of this paper is summarized as follows. Section 2 provides a review of the related work in the field and research background, which is organized in subsections. Section 3 presents the problems addressed in this study, including the hypotheses statements. In Section 4, we describe the behavioural cues that are most significant for the SS of hesitation. The experimental design, the experimental user scenario and additional explanations of the selected aspects of the experimental design that may cause reader confusion are provided in Section 5. Section 6 describes the evaluation methods that were used, while the evaluation results are presented in Section 7. A discussion of the evaluation results is provided in Section 8. Section 9 concludes the study and provides ideas for further work.

2 Research background and related work

In this section, we present the state-of-the-art in the main domains covered in the proposed work to achieve our desired goal. These domains are:

- human–computer interaction,
- social signal processing,
- video-on-demand systems, and
- conversational recommender systems.

2.1 HCI

HCI in its basic form involves the study, planning, and design of interactions between people (users) and computers [17]. HCI is divided into two groups: simple and intelligent HCI [35]. We are interested in intelligent interaction, where the computer understands the meaning of the user's message typically performed using speech and body gestures. Human-centred intelligence (HCI^2) devices [44] are one of the foremost challenges of computer science [35]. The HCI^2 domain bridges the gap between computer science and cognitive science. In the context of HCI^2, computers must have the ability to understand the meaning and context of the information expressed by a user [35]. The newly emerging field of social signal processing (SSP) focuses on this kind of information between human and computer.

2.2 SSP

SSP [45, 54–56] is a research domain that aims to understand social interactions through machine analysis of nonverbal behaviour [55]. SSs are initiated by the human body and present reactions to current social situations. They are expressed with nonverbal behavioural cues that are grouped into five groups [56]: physical appearance; gestures and postures; face and eye behaviour; vocal behaviour; and space and environment. A SS that could be used in HCI is hesitation, which is considered as a kind of uncertainty when a user is faced with a variety of decision choices.

The SS of hesitation is a type of micro-movement called a microslip, which is a nonverbal stutter during the execution of lower level action primitives [38]. Another psychological definition describes hesitation as the elapsing time between the external or internal stimulation of an organism and his, her or its internal or external response [39]. Hesitation can be expressed through facial expressions, head movements, shoulder movements, prosody and special verbal markers such as *eh* or *hm* [24]. However, the 'significant absence' of nonverbal communication can also mean that a person is hesitating.

2.3 VoD systems

The VoD system presents one of the more interesting services in HCI. It enables users to select one video among many. In general, VoD is a system that allows users to select and watch the video of their choice at the time of their choice [19]. The VoD system structure consists of a server for storage of digitized video, a network for transmission of the digital video, and clients that display the video content through a computer or television set. VoD systems can operate in a streaming mode (viewed on the Internet) or in a download mode (the entire program is downloaded and decoded before it can be viewed via a television set). One possibility of how

we could improve the QoE using the VoD is to take the user's SSs into account. These SSs can be used as input information for the conversational RSs that choose the most suitable new video contents, which are then recommended to a user.

2.4 Conversational RSs

RSs are software tools and techniques that predict user preferences and suggest useful items to a user [47]. One of the most important reasons for using RS is to increase user satisfaction when using the system. In conversational RS [47], recommendations are generated based on a natural language dialog between the user and system. The biggest challenges of this domain of RSs are how to design an effective dialogue strategy between the user and system and what actions must be performed in the interactions between them [47]. We believe that using SSs as input information for conversational RSs could help increase the user satisfaction with the system.

2.5 Related work

To the best of our knowledge, there are no attempts to use user SSs in the VoD systems domain. Previous work in the SSP area focused only on analysis of social relations based on voice recognition and was based on corpuses of discussion sessions [20, 53]. However, in our research we recognized SSs that are significantly present with gestures, postures and face or eye behaviour. Few researchers have addressed the problem of action recognition in a context-aware environment [3, 7, 51]. Automatic detection of intra-group interaction is described in [7], while the spontaneous agreement and disagreement recognition approach is presented in [3]. The impact of mimicry on social interaction is shown in [51]. There were also only few attempts of using SSs in HCI [4, 12, 35, 37]. In [4], Branco et al. try to infer the user's expectations regarding task difficulty by watching them just before they start using a photo-copier. In [12], the analysis of user activity during interaction with a photocopier is used to predict the task difficulty. People's hesitant hand motion is proposed as the natural modality for a robot to communicate uncertainty in a human–robot interaction described in [37]. One of the directions on how to use user SSs in HCI is also a system that allows interaction based on electroencephalography (EEG) signals, described in [35]. However, as we can see, the aspect of SSs used in HCI is relatively unexplored. Based on that aspect, we present a new possibility in using SSs in a VoD system, which can improve the interaction QoE.

3 Problem statement and hypotheses

The major problems addressed in this work are how to evaluate the impact of SS of hesitation on user satisfaction while interacting with the VoD system and to show that SS of hesitation is applicable to the VoD system. To the best of our knowledge, there are no similar attempts to evaluate the impact of SSs; therefore, our first addressed subproblem is how to design the experiment. Our next subproblem is the determination of the most significant behavioural cues that best describe hesitation in the context of our VoD system. Based on a study described in [37], we assumed that the SS of hesitation is distinct enough that it can be recognized in HCI. It can be described by the different types of behavioural cues used in [3, 51]. Finally, we address our last subproblem: the evaluation of the impact of SSs. In most cases where systems are evaluated, questionnaires are used [28, 29]. Therefore, we can measure user satisfaction and system usability [29].

3.1 Hypotheses

We tested two hypotheses to confirm or reject the impact of SS of hesitation on user satisfaction while interacting with a VoD system:

- H.1: "*The use of the SS of hesitation in the RS improves the QoE when the user interacts with a VoD system*." and
- H.2: "*The use of the SS of hesitation reduces the content-selection time*."

With H.1, we are trying to determine the link between the SS of hesitation in a HCI and the user's QoE. In our case, the *QoE(u)*, a subjective measure of user experience with the system, is dependent upon two factors (ψ_{SS} and ε_1). We have merged them into the following equation:

$$QoE(u) = \psi_{SS}(SS(u, system)) + \varepsilon_1 \quad (1)$$

where the factor ψ_{SS} represents the impact of SS expressed by the user during an interaction with the system and factor ε_1 presents other possible causes for differences in QoE between the test and control user groups. The theoretical background for Eq. 1 is based on the statistical theory on explained and unexplained variance [36]. Other possible causes may include the impacts of the current mood of the user, current video content on screen, demographics, and user movie expertise. A list of causes measured in our study is presented in Subsection 7.1. Most of them are estimated based on the user answers from the pre-interaction questionnaire.

With H.2, we are trying to determine the link between the SS of hesitation in a HCI and the content-selection time. If SS of hesitation is considered, the user's time to select video content to watch is shorter. Content-selection time is the entire time between the user's first selection (from switching on the VoD system) and the final decision (the user selects video content to watch). However, to test this hypothesis we must ensure the same conditions for test and control user groups. Therefore, we used some assumptions described briefly in Subsection 5.3.

4 When does a user hesitate?

The SS of hesitation is frequently expressed when users interact with and select video items on the VoD system. According to the experimental setup mimicking a real environment, only behavioural cues demonstrated through visual features (such as the speed of a hand movement) are applicable. This is because the system observes the user via a camera. To the best of our knowledge, the SS of hesitation has not been analysed through expressed gestures and behavioural cues. On this basis, we performed a preliminary test where we selected the most significant behavioural cues to describe the SS class {hesitation, no hesitation}. The aim of this test was to find a combination of a small number of behavioural cues that can be used in a decision model with a high recognition rate. These results were then used to model the application in our experimental design through which a human operator recognizes the user SS.

4.1 Methodology for describing the SS of hesitation

Our first task was to observe the users while they interact with our VoD system to obtain valuable information about the behavioural cues for both SS classes {hesitation, no hesitation}. Users interact with gestures to select one of four videos on screen recommended by conversational RS. At each step, the user selects only one video and, based on the selection, four new

videos are displayed. Interaction stops when the user selects a video to watch. For our study, we used seven users (N=7) to interact with the VoD system. We analysed all behavioural cues expressed during user selection of on-screen video content. Altogether, users performed over 30 selections of video items, with more than four video selections per user until a final selection was made. An operator observed the user's selections and noted each behavioural cue made. Each selection was assigned one SS class. Based on observations, we obtained 45 unique behavioural cues for both classes of SS with more than 110 occurrences in total. Obtained behavioural cues included facial, body and arm movements such as a head shake from side to side, a short arm swing away from the body, sucking of the lip, shoulder shrugs, and raising of the eyebrows.

Then we used dummy encoding [22], where only ones and zeroes are used to convey all the necessary information regarding behavioural cue membership. However, hesitation can be also measured by unusual delays in response time; therefore, we included time between two selections in our table of dummy features. We applied a logistic regression model to estimate the absence/presence of selected features:

$$\beta_0 + \beta_1 F_1 + \beta_2 F_2 + \ldots + \beta_m F_m = \ln \frac{p}{1-p}, \tag{2}$$

where β_0 is a constant, β_1, β_2 and β_m are coefficients of the dummy features and F_1, F_2, F_m are selected dummy features. We analysed the data with SPSS software [21].

We tested several other classifiers (Support vector machine, Multilayer perception, Naïve Bayes, and AD tree) and found that logistic regression provides the best accuracy (0.9099), and its balance between precision (0.8824) and recall (0.8333) was good.

We applied known feature selection procedures to narrow down and identify the best from all 45 behavioural cues, but were not able to find a small subset to predict hesitation accurately enough. Therefore, we have manually preselected 12 features that had more than one occurrence and had the largest difference between hesitation and no hesitation classes. After this, we used a forward selection (Wald) method [2] to rank features according to significance and verified the recognition rate for combinations of three, four and five features. On the basis of the selection results, we designed a logit model to decide if a user hesitates or not.

4.2 Results

We used different combinations of features to build the most reliable logit model. Based on the forward-selection method, we obtain the best result by combining the following four features:

- user watching video content, which is then selected for a longer time (F_1),
- user makes a quick gesture when selecting video content (F_2),
- user watches all video content, but none for a longer time (F_3), and
- time between two selections (F_T).

Table 1 shows a confusion matrix for the logit model of the proposed combination of features. We can see that the model has reached a 91 % recognition rate with only four selected features.

Based on the SPSS results, we can present the logit model for the proposed combination of features with the following equation:

$$logit(p) = -5.64 - 5.90F_1 - 1.26F_2 + 1.84F_3 + 0.36f_T, \tag{3}$$

Table 1 Confusion matrix for the logistic regression model where two SS classes are included {hesitation, no hesitation}

		Classified as		
		No hesitation	Hesitation	**Recognition rate (%)**
True class	No hesitation	71	4	94.7
	Hesitation	6	30	83.3
				91

The number denotes the quantity of the classified examples. Overall recognition rate is shaded red

where F_1, F_2 and F_3 are presented with binary values {0 – not expressed, 1 – expressed} and F_T as the absolute time between two selections in seconds. We decided to use binary features because human-operator interface must be manageable to provide real-time responses to the observed user's SSs (the human operator cannot manage the interface and concentrate on the user's interaction at the same time if the complexity of recognized features is too high), and because we simplified the complexity of the experiment to allow for interpretation of results. After we have a model, we can calculate the probability p that tells us if the user is hesitating. If p is equal or greater than 0.5, then user is hesitating, otherwise the user is not hesitating (p less than 0.5).

When the time (F_T) is too long, the impact of this variable on the final result of Eq. 3 is too big in relation to other variables, therefore we must limit it. We replace it by the threshold times from the variables F_1, F_2 and F_3. If time at selected combinations of values F_1, F_2 and F_3 is longer than $meantime + 1.96\ \sigma$ then mean time is used as F_T. Interval $Meantime + 1.96\ \sigma$ represents the confidence interval for normal distribution of time F_T where σ denotes standard deviation of time F_T. The logit model obtained in this study is used in human-operator interface application (see sub-Subsection 5.2.1 for details).

5 Experimental design and experimental user scenario

In this section, we explain the testing scenario and the proposed experimental design for evaluation of the impact of SS on user's QoE in detail. The user's video selection is based on hand gestures and additional information about the produced SS, which can be present through facial expression, hand movements, and eye behaviour. Because automatic gesture recognition does not provide completely reliable results, especially not in real-time, the human operator is used to provide a baseline for real-time action recognition and SS extraction.

We modelled an experimental design, which allows for the control of both factors in Eq. 1 to reliably estimate the contribution of ψ_{SS} to QoE. Its goal was the design of a fair experiment in terms of comparison between the test and control user groups. The test group included users whose SSs during the interaction with the system had been considered. The control group comprised a comparable user group in size and other selected parameters such as age, place of residence, level of education, and sex, whose SSs during the interaction with the system have not been considered. The design also includes some assumptions based on our preliminary work and other studies from other domains used in our work (see Subsection 5.3 for details). We apply an independent-measures experimental design to ensure the feasibility of the experiment and control of the variables.

5.1 Experimental user scenario

The experimental user scenario consists of three events, the first includes activities before interaction, the second presents the interaction between the user and the system, and the third includes the activities after the interaction is done. All scenario descriptions below refer to the test user group.

The first event takes place in room 1 (see Fig. 2) containing the monitor and computer. At the start, the assistant explains the scenario to the user. The user watches the emotionally neutral video (see sub-Subsection 5.3.1) and fills in a pre-interaction questionnaire (see sub-Subsection 6.1.1). After that, the user goes into room 2, where a VoD system using a television screen is installed. The second event starts when the user indicates with a special gesture that he wishes to use the VoD system. The system switches on and the interaction with the system starts.

As part of our system, the RS provides four videos—movie trailers (items) that are projected in parallel on the screen. The user indicates with a gesture, which of the four items he is mostly interested in. The system recognizes how confident the user is about his decision based on the SS of hesitation. If the user is not hesitating, the system displays three similar items (see sub-Subsection 5.3.5) in addition to the selected item. If the user is hesitating, then the system displays four diverse items (see sub-Subsection 5.3.5) according to the current screen. The new items are projected on-screen with sound feedback (see sub-Subsection 5.3.3), which indicates how the system recognized the user's SS. The user repeats the selection process until he finds the item he wants to watch. When the user indicates with a gesture that the final decision has been made (he selects the item he wants to watch), the system expands the selected item (video) to the whole screen and turns on the sound for the video. The user watches the selected item for about 20 seconds (s). After this, the second event is completed. The third event, similar to the first one, takes place in room 1, where the user fills in a post-interaction questionnaire. After he has finished, the scenario is considered complete.

The scenario for the control user group is the same for all three steps. The only difference is how the system provides new items to the user. As we mentioned previously, the control group

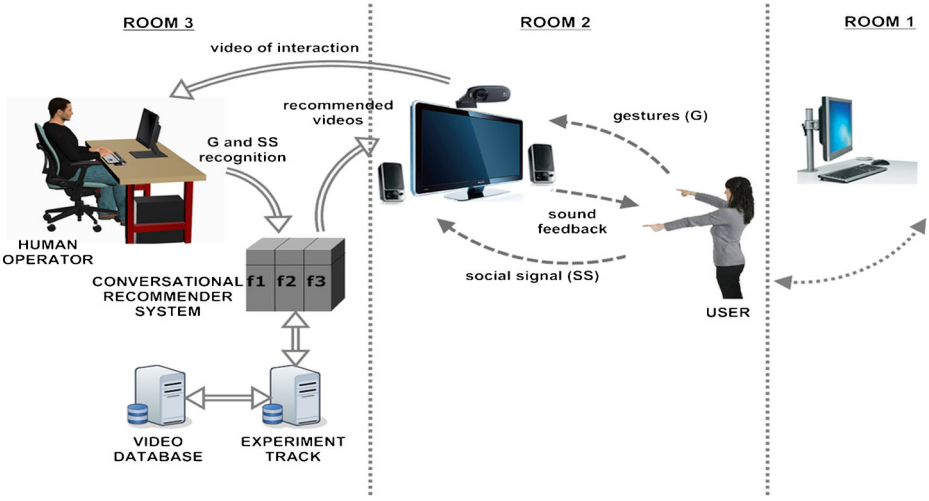

Fig. 2 Experimental environment for the user scenario. Three rooms are needed: one for the human operator, one for user interaction with the system, and one for user activities before and after interacting with the system

user's SS is not considered by the system. On this basis, the system provides three similar items related to the initially selected item. The decision of the system in that case is based only on gestures for video selection without SS. In this particular case, the system is 'dumb' and always gives similar items, in contrast with the system where SS is considered because the system is smarter and understands that sometimes, similar items are not good enough (i.e. the user is hesitating).

Figure 2 shows the experimental environment where the user scenario takes place. It consists of three rooms. Room 1 contained a desktop computer with monitor; the users watch an emotionally neutral video and complete the questionnaires before and after interacting with the system. In our experiment, we used a high-definition (HD) monitor. Room 2 contains the user's interaction system, consisting of a television (tool with which the user interacts), a computer on which the VoD user interface and video database are stored, Internet protocol (IP) camera (video from this camera is used by the human operator), HD camera (for recording user interaction for post-analysis), and speakers (for sound feedback). A human operator sits in room 3, and uses only a desktop computer and monitor on which the human-operator interface is installed. The human operator watches the user interact with the system through the IP camera and makes notes on the actions and SS recognized by the human-operator interface. Based on SS class {hesitation, no hesitation} recognized, the conversational RS recommends new items that are then shown on the user's interface which is accompanied with sound feedback.

5.2 Technical aspect of the experiment

In this subsection, we present the technical details of equipment used in our experiment (see Fig. 2). We used a 32-inch Samsung LCD television. A video database was stored on a HP laptop (ProBook 4720s, Win 7) with 4 GB RAM, and also had the VoD user interface running (with an extended view onto the television as a second screen). The human-operator interface was running on a desktop computer with 2 GB RAM. Video of the user interaction was transmitted through a network with an IP camera (D-Link DCS 3110) connected to the human-operator interface operated by a human operator. A HD camera (Logitech HD) recorded the same scene. These data were stored and used for post-automatic analysis.

As can be seen in Fig. 2, there were two different applications needed for realization of the experiment: human-operator and VoD user interfaces. Both were run as a java file. Communication between applications was based on the HTTP post request method. The same method was used for communication between the VoD user interface and conversational RS. Functions of RS were located on a public server implemented in Python.

5.2.1 Human-operator interface

Current state-of-the-art automatic gesture-recognition algorithms [1] still provide errors in results. In our experiment, this could mean new uncontrolled parameters in an already complex experiment design. This was the main reason why a human operator replaced automatic gesture- and SS-recognition algorithms. Human operator decisions are made in real-time and reported through a human-operator interface.

The human-operator interface (Fig. 3) consisted of various buttons through which the human operator reports their decisions about recognized gestures and SSs. In the middle of the interface is a panel showing live video from the IP camera recording the user. On each side

of this panel are four buttons: two at the top corner and two at the bottom corner. Each group of buttons representing gestures is linked to one video. On the basis of a video selected by the user, the human operator indicates which video is selected and at which step (selection or final selection) using these buttons. Three buttons under the panel (Feature 1, Feature 2, Feature 3) are used for SS recognition. The feature categories are selected based on our preliminary work on SS of hesitation described in Section 4. The human operator indicates at each selection step which of these three features are produced by the user. The binary values of these features {0 – not selected, 1 – selected} and time taken between two selections are used in a logit model (Eq. 3) to calculate the probability that the user is hesitating. Example: *User selects the first video (stretch arm to the top left corner from his point of view). At first, the human operator by pressing the buttons indicates which SS features were produced by the user, and selects the button that indicates selected video. When this button is pressed, the logit model calculates the probability that the user is hesitating and provides a result to the VoD user interface connected to the conversational RS.* The human-operator interface also contains buttons for atom actions and other actions recognition. Atom actions represent hand gestures through which the user communicates with the system. Other actions are defined as actions that cannot be expressed as reactions to current video content, but are defined as a reaction to outside influences. Atom and other actions are part of our further research.

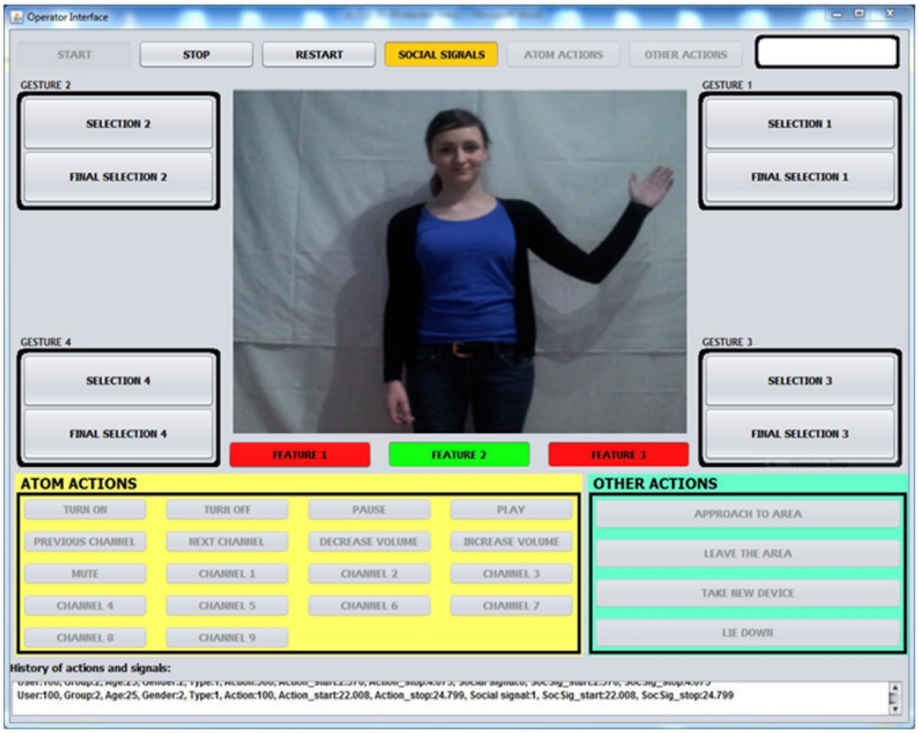

Fig. 3 Human-operator interface. This application is used by the human operator to note all recognized user gestures and behavioural cues that represent a SS class {hesitation, no hesitation}. Feature 1, Feature 2 and Feature 3 buttons are used for SS recognition and selection buttons are used for gesture recognition

5.2.2 VoD user interface

The VoD user interface (Fig. 4) represents the applied version of the VoD system. It consists of four panels where videos are playing. Each decision made by the human operator is provided to the user through a sound feedback interface. The VoD user interface application is connected to conversational RS and the video database. First, it sends information received from the human-operator interface to RS and then uses this response to play videos sourced from the video database. When the user makes their final decision (a gesture that represents the video he wants to watch), the selected video expands to fill the whole screen and the sound of that video starts playing.

5.3 Selected aspects in experimental design

In this subsection, we provide some additional clarification of selected aspects of the experimental design that may be cause reader confusion.

5.3.1 Role of the emotionally neutral video

The initial mood of the users that used our system was not the same. From the aspect of control of our experiment, this was critical. Therefore, we attempted to induce neutral emotions into the users by screening a one-minute video clip at the beginning of the experiment before the

Fig. 4 A VoD user interface consisting of four panels displaying videos

users start interacting with the system. Our assumption, based on previous studies [16, 33], was that the mood state of all users was approximately the same after they watched this video. The video was a documentary clip from National Geographic and portrayed a fish at the Great Barrier Reef [16].

5.3.2 Role of the human operator

Automatic gesture recognition does not provide completely reliable results, especially not in real-time. This is the major reason we introduced a human operator to perform gesture recognition and SS extraction. Based on this assumption, we want to eliminate any new uncontrolled parameters in an already complex design. The human operator provides a baseline action recognition, SS extraction and system feedback to the user in real-time. Because we used simple hand movements, the human operator was not trained to recognize these gestures. However, for SS extraction, the human operator was trained to recognize these behavioural cues selected for our preliminary feature test (see Section 4).

An important factor in HCI systems is responsiveness [49]. In our case system response time (t_{SR}) was measured from the moment a user makes an observable action (gesture of selection) to the moment the user observes a result (new videos on screen). It consists of average human operator response time (t_{Ravg}) and code execution time (t_{CE}). Based on our results t_{Ravg} was approximately 0.7 s and t_{CE} was less than 2 ns. Given that $t_{SR} = t_{Ravg}$ ($t_{Ravg} >> t_{CE}$), is this good enough for the user?

Humans perceive duration based on comparison with expectations established in memory (tolerance threshold) [49]. If the perceived duration is shorter than the tolerance threshold, the user interprets that as fast. Conversely, if the duration is perceived as longer than the tolerance threshold, the user interprets the duration as slow [49]. In our case, where a user interacts with the television and selects the video, the zapping time is the appropriate tolerance threshold. Zapping time [23] is referred to as total duration from the time the viewer presses the channel change button, to the point where the picture of the new channel is displayed on the television. Recommendation ITU-T G.1030 [30] states that for digital IP televisions, the most commonly used televisions currently, channel zapping time should not exceed a limit of 2 s. A zapping time less than 1.4 s is considered to provide a good user QoE [6]. Our system response times met both conditions, therefore we assumed there was no interruption in the user's flow of thoughts and SS expression [41].

5.3.3 System sound feedback

Our next assumption was that the user's emotional response is much less distinctive if he does not know how their SSs and gestures are interpreted than when he knows. It is generally known that behaviour is coherent if the environment is more predictable. An unpredictable environment can lead to an unpleasant user experience and consequently to useless test results [18, 46]. For this reason, we decided on sound for system feedback to the user. We used text-to-speech synthesis for the Slovenian language [25] with predefined sentences. The system played a sound to feedback when the human operator recognizes user gestures or SSs. The text for the test user group included: *"I am offering you four diverse items"*, *"I am offering you three similar items"*, and *"I see you have chosen the item you like"*. On the other hand, the texts for the control group included: *"I am offering you three similar items"* and *"I see you have chosen the item you like"*.

5.3.4 Conversational RS and video database

A conversational RS with no previous knowledge about the user is used. Functions *getInitialItems()*, *getSimilarItems()*, and *getDiverseItems()* (see sub-subsection below) were based on selected videos from the LDOS-CoMoDa research dataset [32] and matrix factorization-based recommender algorithms [31]. However, we did not use all videos from the LDOS-CoMoDa dataset. Our subset contained over 300 videos (movie trailers). All the videos had the same display resolution (632×274) and were in the same multimedia format. The minimum length of a video was 60 s. The distance between movies was computed in a two-dimensional space generated by the first two factors of the matrix-factorization algorithm presented in our previous work [43].

5.3.5 Video selection functions

Based on our testing scenario (see Subsection 5.1), videos were provided to the user according to their produced SS. The VoD system simulates an event in the video rental store or at home. The user wishes to get a video, but he is not sure which one. The support person provides him with four videos (items) and he expresses an opinion. If he hesitates when he selects one item from others, it provides four completely new items. If he is does not hesitate when he selects one item from among others, the selected item remains and three similar items are added. The selection procedure is repeated until a final selection is made. Therefore, we need three video selection functions provided by conversational RS:

$$[hA, hB, hC, hD] = getInitialItems(), \tag{4}$$

$$[hS, hA, hB, hC] = getSimilarItems(hS, h1, h2, h3), \tag{5}$$

$$[hA, hB, hC, hD] = getDiverseItems(h1, h2, h3, h4). \tag{6}$$

Function *getInitialItems* (Eq. 4) provides four videos for the first screen, which cover the whole matrix-factorization space. Function *getSimilarItems* (Eq. 5) provides four videos that are similar to *hS* (selected video); one of them is *hS*, which narrows the search area. Function *getDiverseItems* (Eq. 6) provides four videos that are not similar to *h1*, *h2*, *h3* and *h4*, which expands the search area. The function should diversely cover all factorized video space except those covered by *h1*, *h2*, *h3* and *h4*. The distance metric measuring similarity among movies is based on the matrix-factorization space.

5.3.6 Role of user gestures and SS

In our approach, we used two types of information to manage the system: user gestures and SS. User gestures are used to control the system and behavioural cues (SS) to identify when the user hesitates when selecting the content. Based on the latter, the system expands or narrows the search area. Therefore, only one SS from two classes {hesitation, no hesitation} is transmitted about the content the user sees. The absence of a SS of hesitation means that the user is confident in their decision. In our case, this is the same as when the user does not hesitate. SS is used only to decide between diverse or similar new items. The user uses gesture

to choose one video from others or makes their first decision (selects the VoD system) or final decision (selects the video he wants to watch). Figure 5 presents the set of gestures by which the user selects the VoD system (a), chooses a video (b–e) or makes their final decision (f).

We chose a gesture-based user interface because we wanted to provide natural and intuitive HCI [27], and a way of isolating our experimental parameters (SS). Some advantages of the use of gestural user interface over mouse or remote controller user interface [50] are: (i) it uses equipment we always have on hand; (ii) it can be designed to work from actions that are natural, so there is almost no additional learning; (iii) it lowers cognitive overhead (i.e. there is no interruption of user's flow of thoughts and SS expression); and (iv) it can be used at the most suitable distance (physical space) between the user and the system, where the user feels most comfortable.

From our point of view, the most important advantage is that in a natural environment where gestural interaction is used, conversational activity between a human and the system is much more distinctive; therefore, the SSs regulate the flow of conversation [8].

5.3.7 Data tracked during the experiment

We tracked two types of data during the experiment: user interaction and conversational RS output data. The first type of data includes information about activities and SS of the user recognized by the human operator. Information about input and output videos and selected functions are stored in the second type of data. The whole interaction between user and system is also recorded for the purpose of automatic gesture recognition.

6 Evaluation methods

To evaluate the impact of SS on user satisfaction while interacting with the VoD system, we compared the results of pre- and post-interaction questionnaires for the control and test user

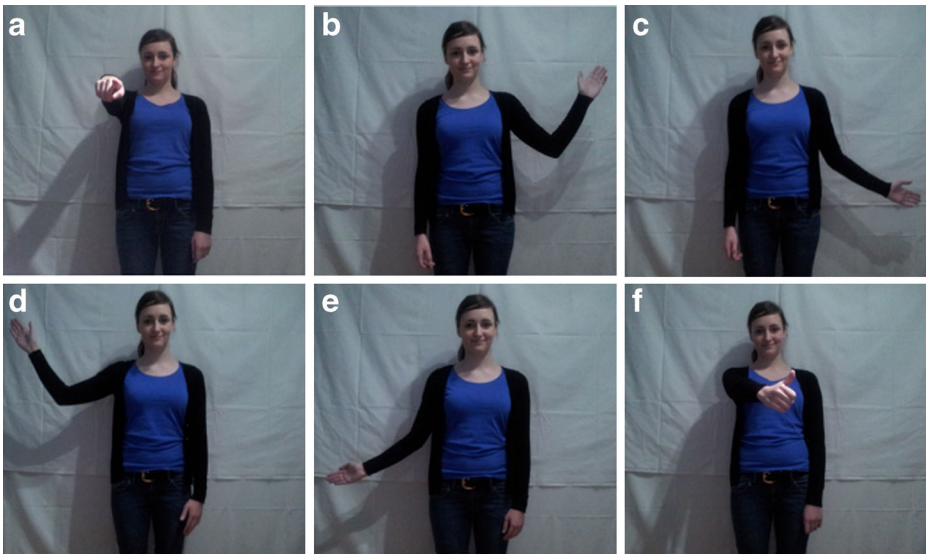

Fig. 5 Set of gestures used by user during interaction. Gestures show the first decision (**a**), selections during interaction (**b**–**e**), and final decision (**f**)

groups. These two user groups are tested to allow realistic effect size estimations of the impact of the SS. Evaluation of hypotheses (see Subsection 3.1) based on statistical tests (Mann–Whitney U test, independent t-test and Fisher's exact test).

6.1 Comparison between test and control user groups

In the test user group, SS induced during interaction with the system is considered. In the control user group, SS is not considered. Our main task was therefore the determination of the size of the impact of SS on user decisions during a communication scenario by comparing the control and test user groups using pre- and post-interaction questionnaires. We measured psychometric characteristics such as reliability and variability.

6.1.1 Pre-interaction questionnaire

The pre-interaction questionnaire comprised 16 statements using a 7-point Likert scale [10, 14] (from completely disagree to completely agree) and one question where only five different replies were possible. It considered the following aspects: user competence about video contents (four statements); user-trusting propensity (four statements); user choice persistence (four statements); user affection towards new technologies (four statements); and possible user pattern preferences (one question). Most of the aspects mentioned above are discussed in [29].

Table 2 shows the statements used in the pre-interaction questionnaire (first part of the table). The psychometric characteristics such as reliability and validity for most aspects were measured. Level of reliability was given by Cronbach's alpha [9]. A value of more than 0.70 was considered acceptable [42]. To establish discriminant validity, there was need for an appropriate average variance extracted (AVE) analysis [15]. The AVE value should be at least 0.50 [15]. Based on values of both characteristics, we eliminated some statements from our analysis (shaded grey in the table). The aspect of user previous preferences is presented with only one statement, therefore psychometric characteristics are not measured.

6.1.2 Post-interaction questionnaire

The post-interaction questionnaire consists of 25 statements and questions using a 7-point Likert scale [10, 14] except in the demographic aspect where different ways were used to collect data. The questionnaire considered the following aspects: user satisfaction with the system (six statements); system usability scale (eight statements); past experiences with similar systems (two statements); user selection time (one statement); user confidence about the accuracy of communication performance (one question); user satisfaction with interpreted SSs (one statement); user satisfaction with recommended videos (one statement); user opinion about task complexity (one question); and personal and demographic information (six questions). The aspects of user satisfaction with system and past experiences with similar systems are described in [29]. In our questionnaire, we used eight of ten statements regarding system usability scale [5] because the system was not complex enough to include statements about inconsistency or integration of subfunctions. How to measure user confidence about the accuracy of communication performance is discussed in [48]. Similar to the pre-interaction questionnaire, all statements can be found in Table 2, which includes psychometric characteristics for most aspects. Some of the statements are eliminated from the analysis because of unacceptable values for reliability and validity (shaded grey in the table). Some of

Table 2 Pre- and post-interaction questionnaire items (first column) used to measure the user's personal characteristics and experience

	Considered aspects	Item	Psych. char.
Pre-interaction questionnaire	Movie expertise	Compared with my peers, I watch a lot of movies. (1)	CA: 0.761 AVE: 0.633
		Compared with my peers, I am an expert on movies. (2)	
		I am a movie lover. (3)	
		I only know a few movies/actors. (4)	
	Disposition to trust people	In general, people really do care about the well-being of others. (5)	CA: 0.786 AVE: 0.658
		The typical person is sincerely concerned about the problems of the others. (6)	
		Most of the time, people care enough to try to be helpful, rather than just looking out for themselves. (7)	
		There are not many people you can really trust. (8)	
	Choice persistence	I am not easily satisfied with a product or service. (9)	
		I waste as little time as possible comparing products/services. (10)	
		When shopping. I have a hard time finding a product that I really love. (11)	
		When I am in the living room watching television, I often cheek other channels to see if something better is playing even if I am satisfied with what I am watching. (12)	
	Familiarity with modern technologies	Compared with my peers, I often use modern technologies or applications (e.g., video-on-demand service, smartphone, Android applications). (13)	CA: 0.813 AVE:0.671
		Compared with my peers, I am an expert on modern technologies or applications (e.g., video-on-demand service, smartphone). (14)	
		I love modern technologies (e.g., smartphone, smart TV). (15)	
		I only know a few modern technologies or applications (e.g. smartphone, smart TV, Android applications). (16)	
	Previous preferences	Imagine a square divided into four equal parts. Which part would you choose? (17)	
Post-interaction questionnaire	User satisfaction	Using the system is annoying. (1)	CA: 0.903 AVE: 0.750
		The system is useful. (2)	
		Using the system makes me happy. (3)	
		Overall, I am satisfied with the system. (4)	
		I would recommend the system to others. (5)	
		I would quickly abandon using this system. (6)	
	System usability	I think that I would like to use this system frequently. (7)	CA: 0.706 AVE: 0.583
		I found the system unnecessarily complex. (8)	
		I thought the system was easy to use. (9)	
		After initial instructions, I think that I would need the support of a technical person to be able to use this system. (10)	
		I would imagine that most people would learn to use this system very quickly. (11)	
		I found the system very cumbersome to use. (12)	
		I felt very confident using the system. (13)	
		I needed to learn a lot of things before I could get going with this system. (14)	
	Previous user experiences	I did not use a similar system. (15)	CA: 0.970 AVE: 0.794
		Use of this system was a completely new experience for me. (16)	
	Perceived selection time	By using the system, I came to the desired content faster. (17)	
	Confidence with completed task	Overall, how confident are you that you completed the task successfully? (18)	
	Perceived SS quality	Overall, the system correctly recognized my desires for choice of video content (the system forwarded the correct feedback sound). (19)	
	Perceived recommendation quality	Overall, I was satisfied with the proposed video contents. (20)	
	Perceived complexity of task	Overall, how difficult or easy did you find this task. (21)	
	User demographic information	Sex (22)	
		Age (23)	
		Place of residence (24)	
		Highest level of education completed (25)	

Items (third column) are divided in different group aspects (second column) that are evaluated with psychometric characteristics (Psych. char.). Those two characteristics were Cronbach's alpha (CA) and AVE. Statements that were dropped from our analysis are shaded grey. Some aspects in questionnaires are presented with only one statement, and therefore were not evaluated. Likewise, some aspects do not reach acceptable CA and AVE values, therefore psychometric characteristics were excluded (diagonal line)

the aspects are presented with only one statement, such as perceived SS quality, therefore psychometric characteristics were not measured.

6.1.3 Statistical analysis

To estimate the data from the questionnaires, we applied the most powerful applicable statistical tests. According to the data type analysed, we used Fisher's exact test, Mann–Whitney *U* test and independent *t*-test for independent samples [26]. An α-value of 0.05 was considered statistically significant, and was chosen because it is commonly and widely used in psychology and social science experiments [13].

7 Evaluation results

The goal of this study was to show that utilization of the SS of hesitation can improve the user QoE when a user is interacting with a VoD system. To confirm or reject the impact of the SS of hesitation, we tested the two hypotheses described in Subsection 3.1. Evaluation of experimental results was based on responses to pre- and post-interaction questionnaires, which were compared between control and test user groups. We used a sample of 28 users (N=28), where control and test user groups contain the same number of users ($N_C = N_T = 14$).

7.1 Potential causes for difference in QoE between control and test user groups

To detect and eliminate the impact of other possible causes for difference in QoE between control and test user groups (see Eq. 1) we compared users according to:

- basic demographic,
- their answers to the pre-interaction questionnaire, and
- video content provided.

In this way, we ensured that the effect on user satisfaction regarding the selected content was caused by the use of the SS of hesitation during the content selection process and not by other differences among users in the control and test user groups. Table 3 presents a list of possible causes for difference in QoE between control and test user groups. We tested a null hypothesis on variables presented in the second column that are divided in five sets (first column). For each set, the statistical analyses are presented with a mean value for control (mean C) and test (mean T) groups, *p*-value and statistical test used.

The first set of possible causes present basic demographics (Tables 4, 5 and 6). We see that there is significant difference (p= 0.031) in age between both user groups. Given that the test user group is almost 6 years older than the control group, we can assume that the test group is not preferred. Based on studies in [34, 40], older users have more difficulty interacting with the system and need more time to complete the task.

The next three sets present the categories from the pre-interaction questionnaire (user movie expertise, user disposition to trust people and user familiarity with modern technologies). There was no significant difference between both groups.

The last set presents video content ratings. All videos in LDOS-CoMoDa dataset are rated (scale 1 to 5) by users that used this dataset. We used these rates to calculate the average of rates of all videos that were recommended to the user. This

Table 3 Table of possible causes for difference in QoE between control and test user groups with results from null hypothesis testing

	Variable	Mean C	Mean T	*p*-value	Test
Basic demographics	Age	28.71	34.43	0.031	*t*-test
	Place			1.000	Fisher T
	Education			0.604	Fisher T
	Sex			0.596	Fisher T
User movie expertise	Pre_Q01	3.36	4.14	0.145	MW *U*
	Pre_Q02	3.43	4.64	0.060	MW *U*
	Pre_Q03	4.50	5.29	0.079	MW *U*
User disposition to trust people	Pre_Q05	5.29	5.29	0.486	MW *U*
	Pre_Q06	4.71	5.14	0.434	MW *U*
User familiarity with modern technologies	Pre_Q13	4.93	5.57	0.522	MW *U*
	Pre_Q14	4.57	5.43	0.194	MW *U*
	Pre_Q15	4.79	5.36	0.424	MW *U*
Video content ratings	Content	3.97	3.75	0.002	MW *U*

The first column presents the set of variables that were measured. Variable names are given in second column, where questions from the pre-interaction questionnaire are labelled with Pre_Qxx (xx denotes unique number of question). The third and fourth columns give the mean variable values for both user groups (mean C – control group, mean T – test group). *P*-values in the fifth column presents the result of null hypothesis testing with a significance level of 0.05. Rows where significant difference was found between groups are shaded red. In the last column, the tests that were used are listed (*t*-test, Mann–Whitney *U* test (MW *U*) or Fisher's exact test (Fisher T)). There are no mean values for place of residence, level of education and sex. Contingency tables were given for all three variables (Tables 4, 5, and 6) based on which *p*-values were measured (Fisher's exact test)

indicates there is significant difference (p=0.002) between both groups. The average of rates was higher for the control user group, indicating that this group was recommended more popular and higher-rated movies. As the control group is not our preferred test group, this difference does not give any advantage in measuring the impact of SS.

Additionally, we also tried to control the current mood of the user. At the beginning of the experiment, neutral emotions are induced by making the user watch an emotionally neutral video (see sub-Subsection 5.3.1). After the user watched this video, the initial mood state of all users was approximately the same.

7.2 The use of the SS of hesitation in the RS improves the QoE when the user interacts with a VoD system

To test our first hypothesis (H.1), we used statements from the post-interaction questionnaire that presented user satisfaction with the system (Table 7). The first tested statement was "The

Table 4 Contingency table

	conG	testG
Female	3	1
Male	11	13

The relationship between user sex and group affiliation. User sample is divided by female and male (variable sex) and by control group (conG) or test group (testG) (variable group)

Table 5 Contingency table

	conG	testG
City	7	8
Suburb	1	1
Village	6	5

The relationship between place of residence, and group affiliation. Place of residence was divided by city, suburb and village (variable place of residence), and users could belong to the control (conG) or test group (testG) (variable group)

system is useful." (Post_Q02) and the second statement was "Overall, I am satisfied with the system." (Post_Q04). Mann–Whitney *U* test was used to measure *p*-value. It is apparent that in both cases there exists significant difference (p_{Q02}=0.022, p_{Q04}= 0.046) between user groups. The mean values for the test group are higher than for the control group. We can conclude that the test user group is more satisfied with the system and found it more useful than the control user group. Likewise, we can accept the null hypothesis for our first hypothesis (H.1), the difference in QoE between a group where the SS of hesitation is considered (test group) and a group where SS of hesitation is not considered (control group).

7.3 The use of the SS of hesitation reduces the content-selection time

To test our second hypothesis (H.2), we used one statement from the post-interaction questionnaire and information about user interaction with the system. Table 8 presents the results of the null hypothesis testing for:

- the statement "By using the system I came to the desired content faster." included in the post questionnaire (Post_Q17),
- the entire time of user interaction with the VoD system in seconds (Int_time),
- the number of user selection gestures during their interaction (Num_ges), and
- average time between two selections in seconds (Avg_Stime).

The Mann–Whitney *U* test and independent *t*-test were used to measure *p*-values. Based on *p*-value results, it is evident that there were no significant differences between groups.

Table 6 Contingency table

	conG	testG
Secondary school	1	3
Tertiary education (professional study)	2	3
Tertiary education (academic study)	5	5
Master degree, doctorate	6	3

The relationship between level of education and group affiliation. Level of education was divided by secondary school, tertiary education (professional study programme), tertiary education (academic study programme) and master's degree or doctorate. The second division divides users by control (conG) and test group (testG)

Table 7 The results of a user satisfaction with the system (QoE measure) as compared between control and test groups

	Variable	Mean C	Mean T	*p*-value
User satisfaction	Post_Q02	4.86	5.64	0.022
	Post_Q04	4.64	5.64	0.046

The null hypothesis was tested using two statements from the post-interaction questionnaire (Post_Q02, Post_Q04). A Mann–Whitney *U* test was used. The results are presented with mean values for both groups (mean C – control group, mean T – test group) and *p*-value (significance level: 0.05). Rows where a significant difference was found between groups are shaded red

Furthermore, the average time of interaction (Int_time) was longer for the test group (mean T) than for the control group (mean C). Similar results were obtained for a number of gestures used in interactions (Num_ges), where on average, users in the test group (mean T) used almost two more gestures than users in the control group (mean C). However, the significant difference between groups for video content ratings and user age may have affected time of interaction and the number of user gestures. The average time between two selections (Avg_Stime) is shorter for the test user group (mean T) than in the control user group (mean C). This might indicate the possible impact of SS, but no statistical significance was found. Use of the SS of hesitation during the interaction did not reduce the content-selection time; therefore, H.2 must be rejected.

We calculated two QoE/time ratios, where QoE was presented as the mean value of the sum of the variables Post_Q02 and Post_Q04 from the post-interaction questionnaire. The first ratio (ratio 1) was calculated between QoE and the entire time of the user interaction with the VoD system (in seconds), while the second ratio (ratio 2) was calculated between QoE and average time between two selections (in seconds). For the control group, ratio 1 was 0.080 and ratio 2 was 0.380; for the test group ratio 1 was 0.078 and ratio 2 was 0.506. The *p*-value results in both cases show that there was no significant difference between both groups for these two ratios. This is expected because the user aim is not being fast in his final decision but to obtain appropriate video content.

8 Discussion

Prior work has documented the importance of SSs in HCI [44, 56]. For example, Ferreira et al. [12] used user SSs produced during an interaction with a photocopier machine to predict the task difficulty. However, these studies have either been theoretical studies [54, 56] without any practical results or have not focused on direct use of SSs in interaction as part of feedback information. In this study, we evaluated the impact of the SS of hesitation on user satisfaction,

Table 8 The results of a reduction in selection time compared between control and test groups

Variable	Mean C	Mean T	*p*-value	Test
Post_Q17	5.00	4.79	0.707	MW *U*
Int_time	77.46	85.80	0.557	*t*-test
Num_ges	5.57	7.36	0.112	*t*-test
Avg_Stime	14.41	11.89	0.151	*t*-test

The null hypothesis was tested for four variables. The results are presented with mean values for both groups and *p*-values with a significance level of 0.05. The final column indicates the test used. The *p*-value column is shaded green as there is no significant difference between groups

when interacting with a VoD system. Two user groups were tested to estimate the impact of SS of hesitation. SS during interaction was considered in the test group, but not in the control group. We compared groups based on responses to pre- and post-interaction questionnaires. We found significant difference between both groups in user satisfaction with the VoD system. These findings extend all previous theoretical studies, confirming that SS of hesitation as additional information in HCI has positive impact on a QoE that reflects user satisfaction with the system. In addition, for a reliable estimation of impact of SS on QoE, we also detected and eliminated other possible impacts on QoE such as basic demographic, user movie expertise, user familiarity with modern technologies, and the impacts of content and mood. This study indicates that the use of SS of hesitation in RS improves the QoE when users interact with a VoD system. To the best of our knowledge, this is the first study to investigate the impact of the SS of hesitation on a VoD system. Our results provide compelling evidence that SS used as input information for RS can improve QoE. However, some other results of this study are worth noting. The average time of interaction is longer for the test user group than for the control user group. We can assume that use of SS of hesitation does not reduce video content-selection time. This could be because of the difference in average user age between groups, where users in the test group are almost 6 years older on average than users in the control group. Older users have more difficulty interacting with the system and need more time to complete the task. Although the average time of interaction is longer for the test group, the time between two selections is shorter for that user group, which may reflect the possible impact of SS. However, this may be because the older users in the test group needed more cycles to complete the task and became impatient. In general, the QoE/time ratios summarise the results of both tested hypotheses. The user's aim in interacting with a VoD system is not hurry to a final decision (selection) but obtaining desired video content and enjoying the user experience. Therefore, users in the test group performed more selections and interacted with the system for a longer time than did those in the control group.

There is also a limitation that must be mentioned such as other potential causes of impact on QoE, e.g., basic demographic, video content ratings. Future work should extend the focus in these causes to eliminate them as impacts on QoE.

9 Conclusion and future work

The use of the SS of hesitation in a VoD system was studied. We modelled an experimental design and an associated experimental user scenario where users use gestures to select among videos on screen. Additional user-produced SS information was used to recommend more suitable new videos in the process of selection. We have identified the most significant behavioural cues that characterize the SS of hesitation. These cues were then used as features for SS of hesitation recognition. Under different assumptions, we have demonstrated that we can measure the impact of the SS of hesitation on user QoE during their interaction with a VoD system when QoE reflects user satisfaction with the system. We have also considered other impacts on QoE (e.g., basic demographic, user familiarity with modern technologies, content ratings). Our finding was that there was a significant difference between the user group where SS is considered (test) and the user group where SS was not considered (control). We expect that these findings can be used to design a VoD system including social context in the user's home environment.

Our future plan is to test new user groups and suggest videos randomly between similar and diverse items. This user group will be compared with a control user group where similar items are always suggested. On the basis of this comparison, we can obtain information regarding

whether our predisposition that 'dumb' system always suggests similar items to the control group is correct.

Acknowledgments Operation partially financed by the European Union, European Social Fund. This work was supported by the EU Seventh Framework Programme FP7 / 2007–2013 through the project PHENICX (grant no. 601166).

References

1. Aggarwall JK, Ryoo MS (2011) Human activity analysis: a review. ACM Comput Surv 43(3). doi:10.1145/1922649.1922649.1922653
2. Bewick V, Cheek L, Ball J (2005) Statistics review 14: logistic regression. Crit Care 9(1):112–118. doi:10.1186/cc3045
3. Bousmalis K, Morency L, Pantic M (2011) Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pp 746–752
4. Branco N, Zagalo N, Branco P, Otero N, Centre A (2011) Blink: observing thin slices of behavior to determine users' expectation towards task difficulty. In Proceedings of the International Conference on Human Factors in Computing Systems - CHI 2011, pp 2299–2304
5. Brooke J (1996) SUS: a "quick and dirty" usability scale. In: Jordan PW et al (eds) Usability evaluation in industry. CRC Press, Taylor and Francis, London
6. Bruzgiene R, Narbutaite L, Adomkus T, Cibulskis R (2013) Subjective and objective MOS evaluation of user's perceived quality assessment for IPTV service: a study of the experimental investigations. Elektronika ir Elektrotechnika 19(7):110–113. doi:10.5755/j01.eee.19.7.5178
7. Carmel M, Kuflik T (2010) Social signal processing: detecting small group interaction in leisure activity. In Proceedings of the 15th international conference on Intelligent user interfaces IUI'10, pp 309–312
8. COMMIT (2011) Sensing for natural interaction. http://www.commit-nl.nl/projects/wp-packages/sensing-for-natural-interaction. Accessed 12 Dec 2013
9. Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. Psychometrika 16(3):297–334. doi:10.1007/BF02310555
10. Dawes JG (2008) Do data characteristics change according to the number of scale points used? An experiment using 5 point, 7 point and 10 point scales. Int J Mark Res 50(1):61–78
11. Diefendorff JM, Richard EM, Gosserand RH (2006) Examination of situational and attitudinal moderators of the hesitation and performance relation. Pers Psychol 59(2):365–393. doi:10.1111/j.1744-6570.2006.00641.x
12. Ferreira JP, Noronha e Sousa M, Branco N, Ferreira MJ, Otero N, Zagalo N, Branco P (2012) Thin slices of interaction: predicting users' task difficulty within 60 sec. In Proceedings of the CHI '12, Extended Abstracts on Human Factors in Computing Systems, pp 171–180
13. Field A (2009) Discovering statistics using SPSS, 3rd edn. SAGE Publications Ltd, London
14. Finstad K (2010) Response interpolation and scale sensitivity: evidence against 5-point scales. JUS 5(3):104–110
15. Fornell C, Larcker DF (1981) Evaluating structural equation models with unobservable variables and measurement error. JMKR 18(1):39–50
16. Gino F, Schweiitzer ME (2008) Blinded by anger or feeling the love: how emotions influence advice taking. J Appl Psychol 93(5):1165–1173. doi:10.1037/0021-9010.93.5.1165
17. Håkansson M (2012) Human-computer interaction. http://www.sics.se/fal/kurser/isd/. Accessed 19 July 2013
18. Hollnagel E, Woods DD (2005) Joint cognitive systems. Foundations of cognitive systems engineering. CRC Press, Taylor & Francis Group, London, p 219
19. Hu A (2001) Video-on-demand broadcasting protocols: a comprehensive study. In Proceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies, pp 508–517
20. Hung H, Chittaranjan G (2010) The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In Proceedings of the international conference on Multimedia, MM '10, pp 879–882

21. IBM Corp. (2012) IBM SPSS statistics for windows, version 21.0. IBM Corp., Armonk
22. IDRE-UCLA (2012) What is dummy coding? http://www.ats.ucla.edu/stat/mult_pkg/faq/general/dummy.htm. Accessed 17 June 2013
23. International Telecommunication Union (2007) Consideration on channel zapping time in IPTV performace monitoring. 4th FG IPTV meeting. Bled, Slovenia, 7–11 May 2007. http://www.itu.int/md/T05-FG.IPTV-C-0545/en. Accessed 15 Dec 2013
24. Jokinen K, Allwood J (2010) Hesitation in intercultural communication: some observations and analyses on interpreting shoulder shrugging. In: Ishida T (ed) Culture and computing: computing and communication for crosscultural interaction. Springer, Berlin
25. Justin T, Pobar M, Ipšić I, Mihelič F, Žibert J (2012) A bilingual HMM-based speech synthesis system for closely related languages. LNCS 7499:543–550. doi:10.1007/978-3-642-32790-2-66
26. Kanji GK (2006) 100 statistical tests. SAGE Publications, London
27. Karam M, Schraefel MC (2005) A taxonomy of gestures in human computer interaction, Faculty of Physical Sciences and Engineering University of Southampton, Southampton. http://eprints.soton.ac.uk/261149/1/GestureTaxonomyJuly21.pdf. Accessed 6 Dec 2013
28. Knijnenburg BP, Kobsa A (2012) Making decisions about privacy: information disclosure in context-aware recommender systems. Institute for Software Research, University of California, Irvine
29. Knijnenburg BP, Rao N, Kobsa A (2012) Experimental materials used in the study on inspectability and control in social recommender systems. Institute for Software Research, University of California, Irvine
30. Kooij R, Ahmed K, Brunnström K (2006) Perceived quality of channel zapping. In Proceedings of 5th IASTED International Conference Communication Systems and Networks, Palma de Mallorca, Spain, 28–30 August 2006, pp 155–158
31. Koren Y (2008) Factorization meets the neighborhood: a multifaceted collaborative filtering model. In Proceedings of the 14th ACM SIGKDD, pp 426–434
32. Košir A, Odić A, Kunaver M, Tkalčič M, Tasič JF (2011) Database for contextual personalization. Elektrotehniški Vestnik 78(5):270–274
33. Lerner JS, Small DA, Loewenstein G (2004) Heart strings and purse strings: carry-over effects of emotion on economic transactions. Psychol Sci 15(5):337–340
34. Leung R, McGrenere J, Graf P (2011) Age-related differences in the initial usability of mobile device icons. Behav Inf Technol 30(5):629–642. doi:10.1080/01449290903171308
35. Lew M, Bakker EM, Sebe N, Huang TS (2007) Human-computer intelligent interaction: a survey. In Proceedings of the 2007 I.E. international conference on Human-computer interaction, pp 1–5
36. Montgomery DC (2009) Design and analysis of experiments. John Wiley & Sons, Hoboken
37. Moon AJ, Panton B, HFM, Van der Loos M, Croft E (2010) Using hesitation gestures for safe and ethical human-robot interaction. In Proceedings of the ICRA 2010, pp 11–13
38. Moon A, Parker CAC, Croft EA, Van der Loos HFM (2011) Did you see it hesitate?—Empirically grounded design of hesitation trajectories for collaborative robots. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp 1994–1999
39. Mu X, Chen Y, Yang J, Jiang J (2010) An improved similarity algorithm based on hesitation degree for user-based collaborative filtering. In: Cai Z et al (eds) Advances in computation and intelligence. Springer-Verlag, Berlin
40. Nicholas D (2005) How age impacts website usability for teens and seniors. http://www.mequoda.com/articles/website-design/how-age-impacts-website-usability-for-teens-and-seniors/. Accessed 20 July 2013
41. Nielsen J (1994) Response times: the three important limits. http://www.useit.com/papers/responsetime.html. Accessed 17 Dec 2013
42. Nunnally JC (1967) Psycohmetric theory, 1st edn. McGraw-Hill, New York
43. Odić A, Tkalčič M, Tasič JF, Košir A (2013) Predicting and detecting the relevant contextual information in a movie-recommender system. Interact Comput 25(1):74–90. doi:10.1093/iwc/iws003
44. Pantic M, Nijholt A, Pentland A, Huang TS (2008) Human-centred intelligent human-computer interaction (HCI^2): how far are we from attaining it? IJAACS 1(2):168–187
45. Pentland A (2007) Social signal processing. IEEE Signal Proc Mag 24(4):108–111. doi:10.1109/msp.2007.4286569

46. Ranne R (2008) Usability and system intelligence. In: Hämäläinen RP, Saarinen E (eds) Systems intelligence: a new lens on human engagement and action. University of Technology, Helsinki, pp 141–157
47. Ricci F et al (eds) (2011) Recommender systems handbook. Springer, New York. doi:10.1007/978-0-387-85820-3
48. Sauro J (2012) Asking the right user experience questions. http://www.measuringusability.com/blog/ux-questions.php. Accessed 19 June 2013
49. Seow SS (2008) Designing and engineering time: the psychology of time perception in software. Addison-Wesley Professional, Boston
50. Song Y, Demerirdjian D, Davis R (2012) Continuous body and hand gesture recognition for natural human-computer interaction. ACM Trans Interact Intell Syst Spec Issue Affect Interact Nat Environ 2(1):11–118. doi:10.1145/2133366.2133371
51. Sun X, Nijholt A, Truong KP, Pantic M (2012) Automatic visual mimicry expression analysis in interpersonal interaction. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR-W'11), Workshop on CVPR for Human Behaviour Analysis, pp 40–46
52. Tkalčič M, Odić A, Košir A, Tasič JF (2013) Affective labelling in a content-based recommender system for images. IEEE Trans Multimedia 15(2):391–400. doi:10.1109/TMM.2012.2229970
53. Vinciarelli A, Dielmann A, Favre S, Salamin H (2009) Canal9: a database of political debates for analysis of social interactions. In Proceedings of the International Conference on Affective Computing and Intelligent Interaction, pp 1–4
54. Vinciarelli A, Pantic M, Bourlard H (2009) Social signal processing: survey of an emerging domain. Image Vision Comput 27(12):1743–1759. doi:10.1016/j.imavis.2008.11.007
55. Vinciarelli A, Slamin H, Pantic M (2009) Social signal processing: understanding social interactions through nonverbal behavior analysis. In Proceedings of the Computer Vision and Pattern Recognition Workshops, pp 42–49
56. Vinciarelli A, Pantic M, Heylen D, Pelachaud C, Poggi I, D'Errico F, Schroeder M (2012) Bridging the gap between social animal and unsocial machine: a survey of social signal processing. IEEE Trans Affect Comput 3(1):69–87. doi:10.1109/t-affc.2011.27

Tomaž Vodlan graduated in the field of electrical engineering at the Faculty of Electrical Engineering, University of Ljubljana, Slovenia in 2010. He joined Agia d.o.o. in 2011 as a young researcher from the industry. His research interests are in the following areas; social signal processing, pattern recognition, machine learning, affective computing and human-computer interaction. Currently he is working on his PhD on processing and utilization of social signals in communication services.

Marko Tkalčič received his PhD degree at the University of Ljubljana in 2011. From 1999 to 2012 he worked as researcher at the University of Ljubljana. From 2013 he is a post-doctoral researcher in the Department of Computational Perception at the Johannes Kepler University in Linz. His research interests include user modelling, recommender systems, affective computing, information retrieval and human-computer interaction. Currently he is investigating how personality and affect can be used to improve various information retrieval methods and human computer interaction. His expertise encompasses also other domains like web applications, machine learning, image processing or human visual perception.

Andrej Košir, PhD, b.s.c math is an associate professor at the Faculty of Electrical Engineering, University of Ljubljana. He is active in a broad research fields including: user modelling and personalization (user models, recommender systems), user interfaces (machine learning method design, statistical analysis), visual data analysis for non-intrusive user data acquisition, optimization, Operations research in telecommunication (optimization of TK systems), social signal processing and statistical analysis of datasets and experiment design. He is a reviewer for projects submitted to TIA technological agency and article submission reviewer in scientific journals.