SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

September 2016, Vol. 59 092106:1–092106:16
doi: 10.1007/s11432-016-5520-1

An exploratory study of multimodal interaction modeling based on neural computation

Lu LU[1,2], Fei LYU[1], Feng TIAN[3]*, Yineng CHEN[1,2], Guozhong DAI[1] & Hongan WANG[1,3]

[1]*Beijing Key Laboratory of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences, Beijing* 100190*, China;*
[2]*University of Chinese Academy of Sciences, Beijing* 100190*, China;*
[3]*State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing* 100190*, China*

Received November 10, 2015; accepted February 26, 2016; published online August 23, 2016

Abstract Multimodal interaction serves an important role in human-computer interaction. In this paper we propose a multimodal interaction model based on the latest cognitive research findings. The proposed model combines two proven neural computations, and helps to reveal the enhancement or depression influence of multimodal presentation upon the corresponding interaction task performance. A set of experiments is designed and conducted within the constraints of the model, which demonstrates the observed performance enhancement and depression effects. Our exploration and the experimental results help to further solve the question about how tactile feedback signal contribute the multimodal interaction efficiency which could provide guidelines for designing the tactile feedback in multimodal interaction.

Keywords human-computer interaction, multimodal integration, interaction model, touch-included interaction, cognition, multisensory integration, neural computation, brain coding

Citation Lu L, Lyu F, Tian F, et al. An exploratory study of multimodal interaction modeling based on neural computation. Sci China Inf Sci, 2016, 59(9): 092106, doi: 10.1007/s11432-016-5520-1

1 Introduction

Multimodal interaction plays an important role in HCI (human-computer interaction) research [1,2]. This paper aims to explore some unnoticed phenomena and rules of cross-modality information influence for facilitating the interaction design on multimodal presentation. Existing work on multimodal interaction mostly focus on improving the computer ability of recognizing human multimodal behavior [1]. However, how to enhance the user understanding of the multimodal information from the computer system is still a field to fill.

Existing work on facilitating people to understand the system-produced information usually lies on the specific areas such as ergonomics [3], visualization [4], interface design [5, 6], etc. Such work mostly focus on visual presentation, while how does the touch-included multimodal presentation influence the corresponding multimodal interaction performance is barely concerned. Touch feedback techniques are

* Corresponding author (email: tianfeng@iscas.ac.cn)

going through a fast blooming era in resent years [7]. Touch experience is enriched and users learn a variety of information from touch feelings [8]. Researchers use touch feedback to enhance the interaction experience [9] or ease the overloading of visual and auditory modality [10]. The new trend of finger touch screen as well as wearable technology promotes touch feedback as a more considerable feedback modality which provide more convenient for HCI [11]. This brought touch modality into the scope of multimodal interaction.

There exist some cross-modality influences which may change the quality of the information received by human [12]. To find out similar influences cross the touch-included modalities may contribute to the touch-included multimodal interaction researches and developments. Existing findings on multimodal presentation influencing information transmission efficiency [13] only consider visual and auditory modality. The difference between touch modality and the visual or auditory modality makes the existing bimodal findings hardly applied on this new multimodal condition.

Some cognitive architectures can help HCI researching and provide modelling tools for human-computer interaction, even including multimodal or multiple-task conditions [14]. However, these architectures are not proposed primarily for addressing HCI problems and the larger endeavor for cognitive studying purposes bring in additional complexities as well as the difficulties of applying them.

Some new findings from neuroscience reveal computational cross-modality influences at the neural response stage [15]. However, the gap from circuits to behavior is really far and the findings are hard to contribute the behavioral-level HCI researches and developments. Intuition tells us the possibility of some similar influence rules can be find in behavioral level and may help interaction design. Yet we are hard to determine how exactly to achieve the potential findings because of the unclearness of experimental constraints and concerned data.

A commentary in the influential journal *Nature Neuroscience* proposes an expected way to bridge the circuits-behavior gap by taking the canonical brain computation rules as an intermediate level between detailed mechanism and overall function [16]. This inspires our work of combining different level of computation rules to model some kind of multimodal interaction tasks to help finding out desired cross-modality influences.

In this paper, we explore the associativity of two proven computation rules to build a multimodal interaction model for touch-included cross-modality influence research. The remainder of this paper begins with a brief overview of the latest findings in neuroscience and cognitive science related with our exploration. Then we propose the model and get the argue of possible influences from the mathematical derivation in Section 3. Section 4 presents the experiments and data analysis guided by the model, and the results reveal the argued phenomena. Finally we conclude the paper with a brief discussion on main contributions and future work in Section 5.

2 Related findings in cognitive sciences

The multisensory nature of perception has drawn attention of much behavioral and neuroscience researchers [17, 18] in recent years. Human brain perceives the physical environment by multiple sensory information such as touch, audition, and vision. These different modalities of information require a robust and coherent percept derived from an efficient merging [19]. This efficient merging of different senses is called multisensory integration. The process will combine information from different sensory systems to influence perception, decisions, and overt behavior [20]. Researchers find that multisensory integration plays a part on enhancing and speeding the detection, localization, and reaction to biologically significant events. New findings are emerged with the latest technologies and views of neuroscience and brain coding [15, 20–22]. Related work shows that, the combination of multimodal information would either enhance [23] or depress [24, 25] the brain performance of information processing. This enlightens our research assumption that, we can find some way for observing similar enhancement or depression effects in behavioral-level multimodal interaction tasks.

2.1 Empirical principles

Researchers in cognitive science are curious about how human recognize the cross-modal stimulus from external environment as well as how the processing mechanism goes. The observable phenomenon and the underlying rules of multisensory integration is revealed as some empirical principles [15, 17] such as assumption of unity [26], spatial/temporal principle of multisensory integration [27, 28], and inverse effectiveness [22].

The assumption of unity proposes that, the observers will be more likely to treat the inputs from two or more sensory modalities as they are originated from a common object or source if the inputs are highly consistent. Therefore, if the information from different modalities share more modal-independent properties, the brain will be more likely to treat them as from a same event [17, 26].

Multisensory integration in cognitive science and neuroscience describes a brain process by which information from different sensory modalities are merged to influence perception, decisions, and observable behavior [20]. As a neural process, the most common assessment of multisensory integration is the consideration of the significant difference between the effectiveness of a cross-modal stimuli combination and its component stimuli for evoking some type of response from the organism [22, 29]. The mentioned response is measured by spikes per second which describes the impulses of neurons. A higher value of spikes per second means a more effective neural response. Therefore operational definitions can be described as follows. When the number of impulses evoked by a cross-modal combination of stimuli is greater than the number evoked by the most effective of these stimuli individually, it shows a phenomenon of multisensory enhancement. Whereas multisensory depression means the less effectiveness of a cross-modal stimuli combination in relate to the most effective individual component stimulus [18, 30, 31]. Multisensory enhancement/depression can represent an increased/decreased likelihood of detecting or initiating a response to the source event of the multimodal signal [18, 22, 30, 31].

The spatial principle of multisensory integration states that the response evoked by a highly effective stimulus from one sensory modality can be suppressed by a less effective stimulus from another modality [15, 27]. The temporal principle of multisensory integration states that multisensory enhancement declines with the asynchrony of the inputs from different modalities and the strongest multisensory integration appears when the inputs are synchronous [15, 28]. The principle of inverse effectiveness states that greater multisensory enhancement is produced by the combination of weak inputs compared to the strong inputs [15, 22].

The research findings from cognitive science and neuroscience reveal the rules of how brain merge the stimuli from different modality, and how the stimuli features alteration influences the merging result. These findings help us understand the characteristics and rules of multisensory integration, which enable us to find some meaningful phenomena for better designing the presentation of multimodal information.

2.2 Computational models

To reduce the uncertainty of the desired potential findings on behavioral-level multimodal interaction, we select two computational models describing the rules of human brain activities under the viewpoint of bridging circuits-behavior gap by canonical computations [16]. The mentioned canonical computations seek to build the mathematical descriptions of the rules revealed by empirical researches. As an intermediate level, these computations guide the research of the underlying circuits and provide formal descriptions for theories of behavior [21]. In this section, the two selected models both have clear mathematical expressions and derivations, and the corresponding case descriptions help to understand the physical meaning of the parameters. We notice the parameter-transference relationship between the two computations, and try to build our new model based on them.

2.2.1 *A normalization model of multisensory integration*

Divisive normalization is a newly proposed canonical neural computation to operate in various neural systems [21]. Since Heeger [32] uses normalization to explain non-linear properties of primary visual

cortex neurons in early 1990s, evidences accumulated to suggest that normalization can explain a wide variety of phenomenon across modalities, brain regions and species such as light adaptation in the retina and mask odorant suppression in the invertebrate olfactory system [21]. Ohshiro et al. [15] propose a divisive normalization model of multisensory integration, which explains the empirical principles of multisensory integration introduced in Subsection 2.1. The model is proposed to provide a simple unified computational tool for understanding and predicting the important neuronal features of multisensory integration.

The divisive normalization model of multisensory integration consists of three layers of neurons (see Figure 1(a) in [15]). Two layers of primary neurons correspond to two different sensory modalities respectively. Each layer is sensitive to one modality inputs. The third layer is composed of multisensory neurons with the assumption that, one single multisensory neuron receives input from a pair of different modality primary neurons with overlapping receptive fields. This assumption is consistent with the research findings in neuroscience [27].

Figure 1(b) in [15] describes the information processing of computing external multimodal stimulus to the multisensory neuron output, and the mathematical representation is given by (1) [15] and (2) [15]. Each unisensory stimulus evokes impulses as the corresponding modality input. The unimodal stimulus is nonlinearly processed to become the third layer inputs, I_1 and I_2, for multisensory integration. After the input nonlinearity, a weighted linear sum is performed by each multisensory neuron as in (2). The unisensory inputs are weighted by modality dominance weights d_1 and d_2. The following expansive power-law nonlinearity expressed by the exponent n represents the transformation from membrane potential to firing rate. At last, the response E is normalized by the normalization pool of all multisensory neurons' response E_j,

$$R = \frac{E^n}{\alpha^n + (\frac{1}{N}) \sum_{j=1}^{N} E_j^n}, \tag{1}$$

where

$$E = d_1 \cdot I_1 + d_2 \cdot I_2. \tag{2}$$

The normalization model of multisensory integration describes the merging process from the external multimodal signal inputs to the integrated output of multisensory integration. The proposed layers and calculations successfully run the model and simulate the neural response of multimodal integration consistent with experimental observations [15]. Meanwhile, the multimodal studies in HCI concern the overt behavior derived from the information recognition result rather than the neural level activity. So we select another computation at a higher level in Subsubsection 2.2.2 to describe a specific behavioral task.

2.2.2 *Bayesian integration in sensorimotor task*

Körding et al. [33] propose a Bayesian integration computation in sensorimotor task. A sensorimotor task means the subject should run some motor control behavior to achieve the task goal, usually with some external and internal multimodal information to support the behavior decision. The computation shows that subjects combine the internal sensory and the external information to get an optimized decision for sensorimotor task performance. It is consistent with the experimental result and can help us to understand and compute the human performance in the sensorimotor tasks. When we perform a sensorimotor task, such as to hit a moving tennis ball, both our sensors and the task have variability. We can only estimate the ball velocity from inadequate information provided by our sensors. The estimation will lead to a hit behavior on specific position. A more accurate estimation leads to a better performance which means a more accurate hit. Multiple modalities of information can be combined to improve the estimation accuracy [34]. This inspires us attending this kind of tasks to study multimodal interaction.

Bayesian integration algorithm has been proved suitable for sensorimotor tasks with motion-adjustment scenarios, which means, users should take adjustment on their motion behavior to achieve some goal [33]. When a task asks for a judgment of object position, people may process the stimulus they received, and

get some estimation. Before the estimation, they should perceive the stimulus to obtain some sensed information with uncertainty. Taking a fat finger pointing as an example, people need to estimate finger point position from the perceived or sensed position which usually is not exactly matching the system-registered position.

Körding et al. [33] describe the relationship between the true information and the sensed ones in a Bayesian formula (see (3)), and assume that human estimate the true information under this Bayes' rule. The variable x_{true} in (3) expresses the true status of the object in the real world, while x_{sensed} expresses the sensed or directly perceived status of the object. In their experimental setup, the x_{true} represents a manipulated shift of the system-produced visual feedback position from the invisible real fingertip position. Subjects need to estimate the shift to adjust their fingertip position for achieving the target position. The optimal estimation $x_{\text{estimated}}$ combines the sensed information and the prior experiences, and this can be represented in (4) when $p(x_{\text{sensed}}|x_{\text{true}}) \sim N(\mu_{\text{sensed}}, \sigma^2_{\text{sensed}})$, and $p(x_{\text{true}}) \sim N(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$.

$$p(x_{\text{true}}|x_{\text{sensed}}) = \frac{p(x_{\text{sensed}}|x_{\text{true}})p(x_{\text{true}})}{p(x_{\text{sensed}})}, \tag{3}$$

$$x_{\text{estimated}} = \frac{\sigma^2_{\text{sensed}}}{\sigma^2_{\text{sensed}} + \sigma^2_{\text{prior}}}\mu_{\text{prior}} + \frac{\sigma^2_{\text{prior}}}{\sigma^2_{\text{sensed}} + \sigma^2_{\text{prior}}}\mu_{\text{sensed}}. \tag{4}$$

The Bayesian integration process is to combine the external conditions and internal experiences to get estimation on the real object status, so that the person can make the right decision on body movement for the object operation. The research manipulate the reliability of the multimodal feedback signals in their experiment to avoid the influence by uncontrollable variables in the process of human information perception. This setting improves the accessibility and controllability of the sensed information x_{sensed}, which contributes the human estimation as an input variable.

3 Multimodal interaction model based on neural computation

We can build a new computable relationship based on the two computations introduced in Subsection 2.2 since they are both considering the perceived information from external presented sensory stimulus. The normalization model [15] computes the multisensory integration without considering the cognitive level information processing, and the Bayesian integration [33] presents the human-brain combine manner of sensed information and the other task-related conditions, with the manipulation of the sensed infromation. To be specific, as shown in Figure 1 in [33], sensed information can be represented as some kind of information distribution related to the stimulus presentation. And when this distribution follows the Gaussian distribution, it will match the (4). Multimodal stimulus can also produce such sensed information about the object state, and the resulted information can be processed as one of the parameters of the Bayesian integration [33] as the $p(x_{\text{sensed}}|x_{\text{true}})$ in (3). In the normalization model [15], although the output of the multisensory integration represents the neural response, it can be corresponded to such human concerned information by the third layer neuron setup. This relationship of parameter transfer makes it feasible of building a combined model to represent how different multimodal information presentation would influence the sensorimotor task performance.

3.1 Modeling analysis: human processing on multimodal information

There shows a black box in Figure 1(a). We argue that, some kinds of cross-modal influences contribute to the final task performance, and the phenomena can be observed to show some influencing rules which may help multimodal interface design. The relationship of parameter transfer between the two models mentioned in Subsection 2.2 helps us to disambiguate the black box into a grey box as in Figure 1(b). From the view of cognitive psychology [35], we can see the procedure as in three stages of perception-cognition-behavior procedure shown in Figure 1(c).

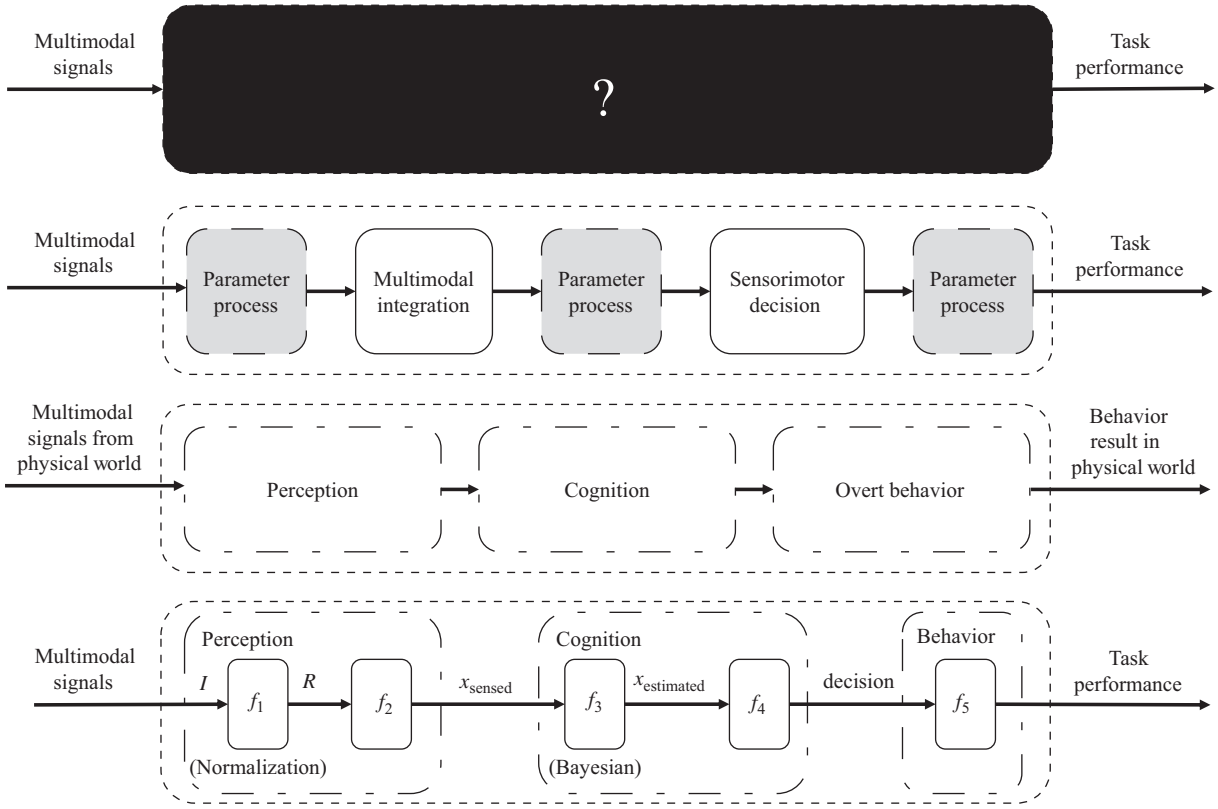

Figure 1 Modeling analysis: human processing on multimodal information. (a) Black box; (b) grey box; (c) three stages; (d) five mappings.

From the intermediate level computation, which is the Bayesian integration [33], we notice that the feedback of user fingertip position can influence the position adjustment. Their experiment [33] manipulates the feedback to eliminate the ambiguity of the sensed information. The manipulated feedback can be considered as some kind of amplified uncertainty of the sensed information. In our study scope, we replace this manipulated uncertainty by the result of human multisensory integration, i.e., different multisensory feedback situation leads to different level of sensed information uncertainty. Take the touchscreen interaction for example, user estimation of the fingertip position has some uncertainty because the fingertip covers a bigger region than the system-registered position. Presenting some kind of feedback signal for the fingertip position may reduce this uncertainty. Refer to the multisensory integration rules mentioned in Section 2, different multimodal feedback presentation may conclude different level of sensed information uncertainty which corresponds to different level of neuronal response. To further elucidate the procedure, we analyze it as several mappings (see Figure 1(d)) reflect the discussion before:

Mapping 1. Multimodal feedback signals provided by the system → Multisensory integration result of multimodal signals;

Mapping 2. Multisensory integration result of multimodal signals → Sensed task object state;

Mapping 3. Sensed task object state → Estimated result of the real-world task object state;

Mapping 4. Estimated result of the real-world task object state → Motion behavior decision for the task;

Mapping 5. Motion behavior decision for the task → Task performance.

Mappings 1 and 2 reflect the perceiving procedure described in the normalization model [15], Mappings 3 and 4 reflect the cognitive procedure of the sensorimotor task described in the Bayesian integration [33], and Mapping 5 refers to the behavior procedure that user actions lead to the task execution result.

Mapping 1 can be described as a simple normalization calculation in (5) which corresponds to the normalization model introduced in Subsubsection 2.2.1. Notice that, the layered neuronal model described in [33] can be modified to represent different physical meanings by modifying the neuron setup. So in our

setup, R_j in (5) corresponds to the output of Mapping 1, where $j \in \{1, 2, \ldots, k\}$ corresponds to a specific recognition result conclusion$_j$ within all possible alternative conclusions concluded from the multimodal stimuli $I = \{I_1, I_2, \ldots, I_m\}$ as the input condition set, where I_m represents the unisensory input from modality m. T_j presents the tendency of the input condition set supporting conclusion$_j$, and the value is calculated by (6) as a weighted linear sum of each unisensory inputs. Note that different alternative conclusion$_j$ will have different support tendency from the same input condition set. Accordingly, T_k is the tendency of the input condition set supporting conclusion$_k$, and all possible alternative conclusion constitutes the normalization pool represented by the sum $\sum_k T_k^n$ (see Figure 1(b) in [15]). This normalization pool shows that, the support tendency of the other alternative conclusion will weaken the supportiveness of conclusion$_j$ which matches the real world observation [15, 21]. The calculation result of this equation shows the support level of the condition set to a particular conclusion$_j$. The output of Mapping 1 should include the support level of every alternative conclusion, that is, $R = \{R_1, R_2, \ldots, R_k\}$. For each different scenario setup, the corresponding parameters γ, α and n are constants [21] in (5), and the parameter γ makes sure that $\sum_k R_k = 1$,

$$R_j = \gamma \frac{T_j^n}{\alpha^n + \sum_k T_k^n}, \tag{5}$$

where

$$T_j = \sum_m d_{jm} I_m. \tag{6}$$

Mapping 2 needs to transform the Mapping 1 output into the Mapping 3 input. This transforming can be represented as $p(x_{\text{sensed}}|x_{\text{true}}) = f_2(R)$. The transformation can be analogically explained as in [33], where the different kind of manipulated feedbacks lead to different probability distribution of possible sensed information. For example, if the sensed task object status can be correspondingly approximated by Gaussian distribution, then it can be represented as $p(x_{\text{sensed}}|x_{\text{true}}) \sim N(\mu_{\text{sensed}}, \sigma^2_{\text{sensed}})$, where μ_{sensed} and σ_{sensed} correspond to the conclusion distribution of Mapping 1 output R. Under such conditions, we define the mean of R represented distribution as the conclusion$_j$, where j meets the max value of $R_j \in R$. Specifically, if there exists more than one fitted j, choose the median value of the conclusion$_j$ set, i.e., $\mu_R = \text{Mid}\{\text{conclusion}_j|_{\text{Max}\{R_j \in R\}}\}$. Since $\sum_k R_k = 1$, according to the properties of discrete random variables, we define the variance from $\sigma_R^2 = \sum_{j=1}^k (|\text{conclusion}_j - \mu_R|^2 \cdot R_j)$ where $|\text{conclusion}_j - \mu_R|$ represents the difference between conclusion$_j$ and μ_R. Then we can get $\mu_{\text{sensed}} = g_1(\mu_R)$ and $\sigma^2_{\text{sensed}} = g_2(\sigma_R^2)$.

Mapping 3 corresponds to the procedure described in (3). The users get the estimation from the sensed information combined with their existing prior experiences. If the sensed information follows Gaussian distribution, then the procedure can be represented by (4). The subscript prior corresponds to the known measurement of information uncertainty, and corresponding parameters can be analyzed from the specific case settings. Take the experimental setup in [33] as an example. The subjects learn the prior distribution by repeatedly doing the task with final position feedback, and the learned prior distribution readjust the sensed information to get an optimized estimated information. Therefor the different level of sensed information uncertainty will lead to different estimated information (see Figure 1 in [33]). This Mapping 3 represents the high level cognitive correction on the primary recognized result, which procedure integrates the external and internal factors related to the task performence.

Mapping 4 can be represented as decision $= f_4(\text{goal}, x_{\text{estimated}})$. Take a drag task as an example, the task goal is to drag the target to a specified position, thus the drag direction as the motion behavior decision in Mapping 4 can be derived from the estimated target position in Mapping 3 outputs.

Mapping 5 represents the causality from the motion behavior decision to the task performance. This can be represented as performance $= f_5(\text{decision})$. Some simple task may demand only one decision, the corresponding behavior to achieve the task goal will lead to the task performance. However, in most tasks like racing games or parkour games, users may repeat the decision-making cycle and adjust their motion behavior to move the target to avoid new risks of damage, thus the task performance is the result of a series of decision-making and corresponding behavior cycles, which correspond to the whole

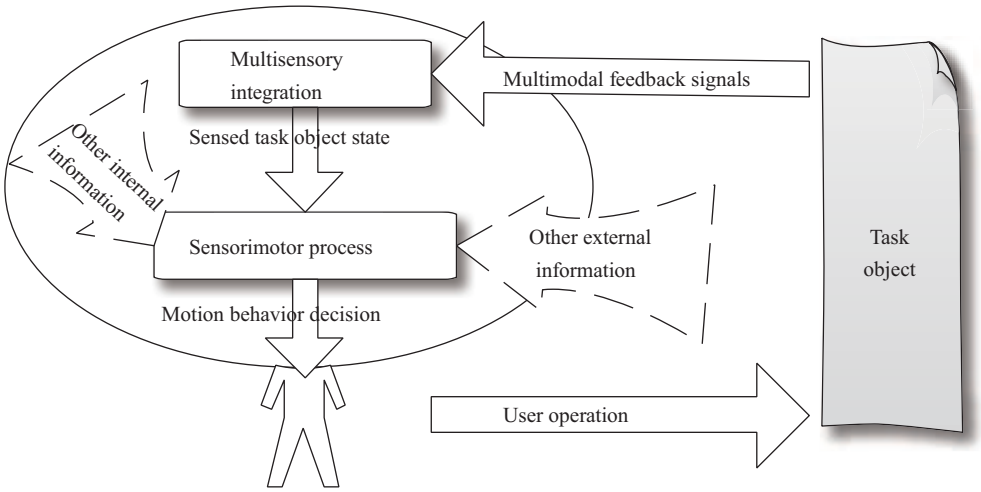

Figure 2 Multimodal interaction model based on neural computation.

procedure from Mapping 1 to Mapping 4 as one of the cycles. This implies some specialty of Mapping 5, so we propose some constraints to avoid excessive complexity. We assume that every cycle is dealing with the same configuration of information and the outputs of each cycle contribute the task performance equivalently. In this way, a positive correlation is established between the quality of human motion behavior decision and the task performance, the more accurate the decisions, the higher performance the task.

3.2 A multimodal interaction model based on neural computation

We propose a combined model based on the analysis above. The proposed model uses the normalization model to describe the multimodal integration procedure, and the Bayesian integration method to describe the interaction task decision-making procedure. The task performance is derived from the quality of the decision-making results. This whole process completes the derivation from the multimodal presentation to the derived task performance, and can get meaningful arguements through the derivation.

Applicable scenario constraints. The multimodal interaction model based on neural computation proposed in this paper is for the scenario of sensorimotor tasks where healthy users should make the behavior decision to change the task object state. Specifically, when the task goal needs more than one decision cycle to achieve, the decision cycles should have the same input and output set with different values, and the outputs should contribute the task performance equally.

Modeling target. The proposed model mainly describes the procedure from the multimodal signals to the corresponding task performance, which includes recognizing information from external signals, estimating the target state from the recognized information, making behavior decisions from the estimated state, doing the behavior to achieve the task goal, and finally getting the task performance from the task completion result. The modelling aims to reveal the influence of different multimodal presentation on the corresponding task performance.

Composition and description of the model. The composition of the model and the corresponding information processing is shown in Figure 2. The user processes all external and internal information to get the decision for task operation. Our concerned information is the multimodal feedback signals, this can be processed by the multisensory integration procedure [15] to get the sensed task object state. Then the sensorimotor process represented by the Bayesian integration model [33] considers the sensed state and other external and internal information to make the behavioral decisions. The behavioral operations on the task object lead to the final task performance. The mathematical description can be given follow the analysis in Subsection 3.1:

Mapping 1. $R = f_1(I)$, where $R = \{R_1, R_2, \ldots, R_k\}$, $I = \{I_1, I_2, \ldots, I_m\}$, and each R_j is calculated from I by Eqs. (5) and (6);

Mapping 2. $p(x_{\text{sensed}}|x_{\text{true}}) = f_2(R)$, the Gaussian distribution of the sensed information can be represented as $p(x_{\text{sensed}}|x_{\text{true}}) \sim N(\mu_{\text{sensed}}, \sigma^2_{\text{sensed}})$, where $\mu_{\text{sensed}} = g_1(\text{Mid}\{\text{conclusion}_j|_{\text{Max}\{R_j \in R\}}\})$, $\sigma^2_{\text{sensed}} = g_2(\sum_{j=1}^{k}(|\text{conclusion}_j - \mu_R|^2 \cdot R_j))$;

Mapping 3. $x_{\text{estimated}} = f_3(p(x_{\text{sensed}}|x_{\text{true}}), p(x_{\text{true}}))$, for Gaussian distribution,

$$x_{\text{estimated}} = \frac{\sigma^2_{\text{sensed}}}{\sigma^2_{\text{sensed}} + \sigma^2_{\text{prior}}}\mu_{\text{prior}} + \frac{\sigma^2_{\text{prior}}}{\sigma^2_{\text{sensed}} + \sigma^2_{\text{prior}}}\mu_{\text{sensed}};$$

Mapping 4. $\text{decision} = f_4(\text{goal}, x_{\text{estimated}})$;

Mapping 5. $\text{performance} = f_5(\text{decision})$.

3.3 The influence between multimodal presentation and the interaction task performance

In the model we proposed, the multimodal presentation determines the perceived information quality and then leads to corresponding decisions for task performance. According to the empirical principles of multimodal integration correspond to Mapping 1, a better presentation will lead to multisensory enhancement while an inferior presentation will lead to multisensory depression.

The principles about multisensory enhancement are introduced in Section 2, they describe the rules of how to get multisensory enhancement/depression through different multimodal presentations. As discussed in this section, optimizing the presentation will enhance the stimulus perceiving so that more reliable information can be perceived for estimating the target state, thus the task performance can be improved without any other condition change. On the contrary, a less efficacy presentation will reduce the task performance because of the increased information uncertainty.

The calculation in Mapping 1 can simulate the spatial/temporal principle and the inverse effectiveness. The spatial principle describes two single modality stimuli presented together at the same time, and states out the relationship between the signal presentation and the multisensory enhancement/depression. We have already stated that multisensory enhancement may cause a better task performance while multisensory depression may lead to a reduced task performance. Therefore we can deduce some likely effects of multimodal interaction task performance due to multisensory integration and design experiments for observable evidences. The temporal principle suggests that synchronous signals from different modalities can lead to the strongest multisensory integration. Coincidentally, without explicitly using this principal, existing design work always present the multimodal signals at the same time if they consider the same object. The inverse effectiveness discusses the stimulus intensity difference at the neural level. Although this difference may become hard to identify at the observable behavioral level, it can suggest our experimental design to try to focus on the subtle information so that a more observable enhancement effect can be looking forward to.

According to the spatial principle, a lower effective stimulus from one modality will depress the response evoked by a highly effective stimulus from another modality. That is, if two unimodal stimuli from different modality have a large difference of intensity, it is more likely to have a multisensory depression when combine them together at the same time. On the other hand, if the two share a similar intensity, it is more likely to observe a multisensory enhancement [15]. This can lead to the argument that, when the user receives two unimodal stimuli from different modality at the same time, there may exist below effects:

Argued performance depression effect. If an individual unimodal signal can support a much better task performance than the other modality, the two may get a reduced task performance than the better individual one due to the reduced information quality because of the multisensory depression.

Argued performance enhancement effect. If an individual unimodal signal can support a similar task performance as the other modality, the two may get a much better task performance benefited from the more valid information thanks to the multisensory enhancement.

4 Experiment

We preliminarily designed a set of experiments in order to find observable evidences of the argued effects. The experiment follows the applicable scenario constraints of our model mentioned in Subsection 3.2. The constraints ensure a higher possibility of observing desired phenomena because of the variable influences within the model. Our experiment requires a fingertip line tracing task where the subjects need to adjust their fingertip position within many sub cycles described in Mappings 1–4. The sub cycles contribute a number of data to each one short task. This makes the experiment to be simple and reliable.

4.1 Participants and apparatus

Twelve subjects participated in the experiment, 8 males and 4 females, age from 24 to 33. All of the participants are right-handed, and use their right index finger to operate the experiment task.

The experiment app is implemented in MI3 mobile phone, and the tactile feedback technology is provided by the TouchSense tactile feedback technology from Immersion Corporation.

4.2 Task and procedure

As shown in Figure 3(a), we designed a fingertip line tracing task. The participants are required to slide their fingertip through the middle of the line shown on the screen, from the top-left white dot center to the bottom-right white dot center. If the fingertip deviates 1 mm from the middle of the gray line (i.e., the fingertip lies beyond the scope of the central light gray line), the evaluation of task performance will reduce. When the fingertip deviates 2 mm from the gray line (that is, the fingertip lies beyond the outside boundary of the peripheral dark gray line), the task will be evaluated as a failure. Under this setup, it is nearly impossible for participants to track the line precisely since the estimation of the exact fingertip position is hard while the allowed deviation is quite small and the performance is influenced by some factors like the fingertip covering, the operational instability, etc. This task contains several decision-making cycles of fingertip movement adjustment, that is, the user should keep adjusting their fingertip position to ensure it sliding right through the line center. The whole task performance depends on the performance of every sub cycle.

Based on this setup, our experiment is designed to test several different single modal and multimodal signals for improving the corresponding task performance. The design goal of each signal is to help users catch the exact position of their fingertip in real time, to help the decision-making of the adjustment of their fingertip movement and to improve the task performance. The experiment includes signals of a single visual feedback, a single tactile feedback, and a combined feedback of both visual and tactile modality. Visual modality signals play an important role in HCI, making it the first choice in the study of multimodal interaction. Despite that tactile interaction is rising to become a new concerned study topic and spreading through various applied devices, the contribution to the interaction efficiency by tactile signals is still being questioned. Therefore, we chose the visual modality and the tactile modality as the research targets. Current tactile design researches provide evaluation on subjective experience like realism and satisfaction, while the contribution of tactile signals to observable multimodal interaction performance is usually neglected. We expect this experiment to help finding the influence rules of how the three different feedback treatments, which are respectively corresponded to different multimodal presentations, influence the task performance, meanwhile the role of tactile feedback in interaction design can be revealed clearer. Thus, the experiment includes four different feedback treatments: treatment A refers to no additive feedback and only deploys the basic setup, treatment B refers to visual feedback upon the basic setup, treatment C refers to tactile feedback upon the basic setup, and treatment D refers to a combined feedback of both visual and tactile modality upon the basic setup.

Treatment A follows the basic setup described in Figure 3(a). The goal of the experiment is to compare task performance under the four different feedback treatments, and to find the observable evidences for the argued influencing effects. So we set treatment A as a reference, to measure the influence of different treatments to the task performance. The other treatments respectively add different feedback signal to

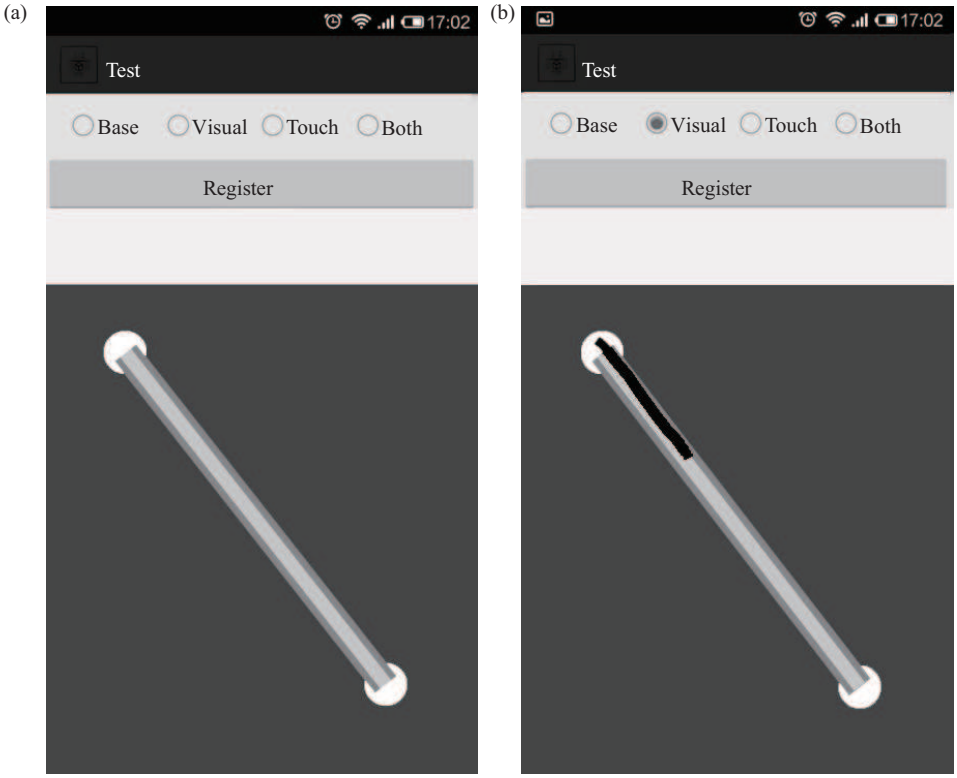

Figure 3 Interface of the experiment app. (a) Basic setup; (b) visual feedback setup.

this baseline treatment A. This setup excludes the other influence factors, and only considers the feedback signal design.

Treatment B is shown as Figure 3(b). A visual stroke will be shown with the movement of fingertip sliding over the touch screen, so that the user can see their fingertip track. The fingertip position and its deviation from the gray line center can be estimated by comparing the stroke with the gray area as reference. This estimation will help the user adjusting the finger position and movement direction to achieve a better performance, which is, to get closer to the middle of the gray line. Whereas this estimation is not accurate for two reasons. The first reason lies in the limitation of human physical ability. Human eyesight can hardly detect small differences on millimeter level, especially the slightly difference of whether the fingertip stroke out of bounds. The second reason is because of the fingertip covering. The fingertip range will cover the corresponding part of the screen. This brings difficulties for estimating the fingertip position in real time and makes the movement adjustment harder.

Treatment C provides tactile feedback. Two kinds of tactile signals are provided by the hand phone vibrator. Since the fingertip moved out of light gray area, tactile signal 1 will be provided. When the fingertip moved beyond the dark gray periphery, tactile signal 2 will be actuated. In this way, users can feel whether their fingertip is out of bounds. However, due to the human eyesight limitation and the fingertip covering, it is still difficult to get accurate estimation of fingertip position and to achieve perfect decision of movement direction.

Treatment D provides both visual and tactile feedbacks as a combination. Similar to the other treatments, the problem of human physical limitation and fingertip covering still exists. Users cannot get the precise fingertip position and perfect movement direction.

4.3 Design

A within-subject factorial design was adopted. The order of different treatments was counterbalanced across participants. To increase the reliability of the data, each participant was asked to complete the task 6 times for each treatment. In total, we collected data from 288 trials by 12 participants (4 treatments $\times$

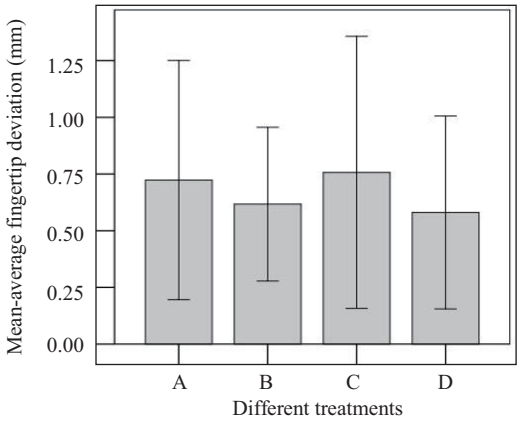

Figure 4 Average fingertip deviations in each treatment.

6 repetitions × 12 participants). For each trial, we collected data about fingertip deviation (the distance between every fingertip tracing point and the task line center) and time cost. Before the experiment began, each participant had 10 min to practice until they feel familiar with the setup. The experiment lasted approximately 5 min for each participant. Participants could have 10 s to break between different treatments.

4.4 Results and discussion

4.4.1 *General result*

We firstly compared the task performance under the four different feedback treatments in common ways. Figure 4 demonstrates the distribution of the average fingertip deviation from the middle line under different treatments for all participants. The average fingertip deviation is the mean of every sub cycle fingertip deviation in one task completion. A lower deviation value means a better adjustment decision making. And a lower average fingertip deviation value refers to a better task performance.

Significant differences in the general comparison show that, treatments B and D are significantly better than treatments A and C. The lower average fingertip deviation means a better performance under which the fingertip is closer to the middle line while the task operation.

This result confirms the existing research view that, under certain setup, visual signals will be more dominant than tactile signals for improving the interactive task performance, and the tactile signals usually contribute little for this.

4.4.2 *Difference between modalities*

Since our research focus is to find the influence of cross modal differences upon the bimodal feedback task performance, we further analyzed the experiment data.

Besides taking the average fingertip deviation as the measurement, we calculate the mean average fingertip deviation of each user under each treatment respectively. Furthermore, we calculate the difference of the mean average fingertip deviation between treatments B and C for every participant. This difference represents the cross modal difference we attend to. That is, the different influences on task performance by different unimodal feedbacks.

Based on this measurement of cross-modal difference, we reclassify the data from treatments B and C. By experimental setup, treatment B represents visual modality feedback while treatment C represents tactile modality feedback. Nevertheless, we notice that, the influence effects argued in Subsection 3.3 have no concern about specific sensory modality. Instead, they concern about advantaged or disadvantaged modality which would contribute the task performance better or worse. For some users, individual visual feedback will get better task performance than individual tactile feedback dose, and for the other users,

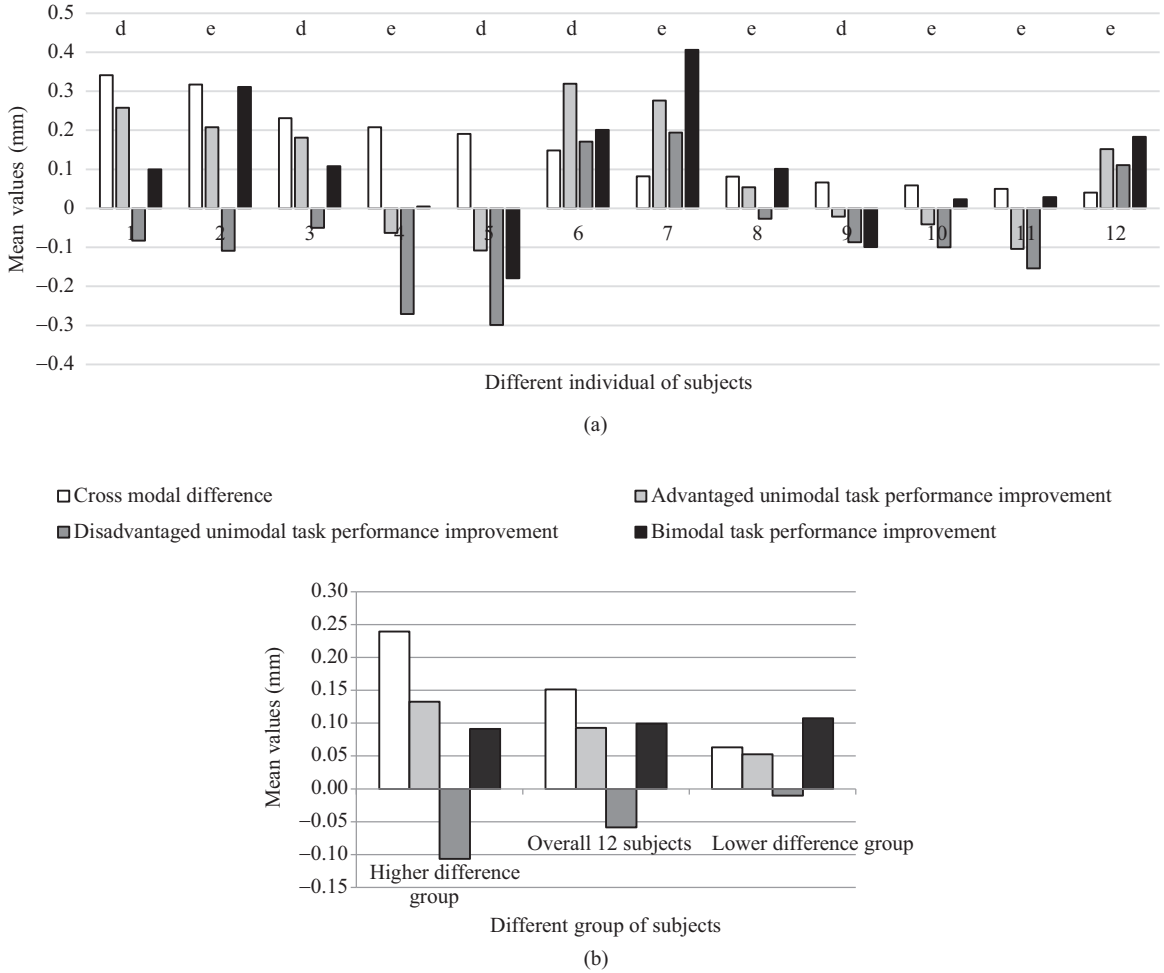

Figure 5 The influence of different feedback treatment consider the advantaged or disadvantaged modality. (a) Individual view; (b) grouped view.

touch will get the advantage. So we compare the cross-modal difference described before and reclassify the data of treatments B and C into advantaged modality and disadvantaged modality.

We order the individual cross-modal differences between advantaged modality and disadvantaged modality in Figure 5(a), and present the cross-modal difference as well as the performance improvement compared with treatment A of the advantaged-modal , the disadvantaged-modal and the bimodal situation for each individual. The performance improvement compared with treatment A demonstrates the different promotion of task performance due to different feedback treatments. This comparison limits the value range and allows negative values. In Figure 5(a), the value of cross-modal difference is represented as the white bar, and arranged from large to small within the sequence of subject 1 to subject 12. The light grey bar beside the white bar represents the advantaged-modal performance promotion. And the black bar represents the bimodal performance promotion. If the black bar is higher than the light grey bar, we call it a performance enhancement for the individual, noted e in Figure 5(a) for each subject. And if it is lower, we call it a performance depression, noted d in the figure. We notice that subject 10 to subject 12 all match the argued performance enhancement effect. And from subject 7 to subject 12, there shows only one depression case. While the argued performance depression effect is not obviously shown in this Figure 5(a).

Based on this observation, we further explore the correlation between the performance enhancement value and the cross-modal difference. we calculated the difference between the bimodal task performance improvement and the advantaged unimodal task performance improvement as a new value, performance enhancement, for each individual subject. This new value allows negative situation. Two-step cluster

analysis is used to reveal naturally occurring subgroups of this new value as well as the cross-modal difference value. The performance enhancement values are clustered into two groups of performance enhancement group and performance depression group, respectively correspond to the positive and negative values (silhouette = 0.8). And the cross-modal difference values are clustered into two groups of higher difference group and lower difference group, each has six members (silhouette = 0.8). Based on these naturally clustered enhancement evaluation and the difference level, we run a one-tailed Spearman's rank correlation test, and conclude a significant negative correlation at the statistical significance level of 0.05 ($\rho = -0.507$, $p = 0.046$).

In order to get a more intuitive view, we compare the grouped mean values of the higher or lower difference groups, and draw Figure 5(b). The mean values of each group show that, higher difference group gets a weaker task performance in bimodal condition than in the advantaged-modal condition, whereas lower difference group gets a better task performance in bimodal condition than in the advantaged-modal condition. The overall mean values of the all 12 subjects are demonstrated between the group mean values as the higher difference group values demonstrate on the left and the lower difference group values demonstrate on the right of the graph in Figure 5(b). The overall mean values show a similar pattern as in Figure 4, even though the unimodal signals here is reclassified as advantaged or disadvantaged modality. The pattern of the three groups changes with the hight of the white bar. This shows the cross-modal difference as a key influencing factor on the bimodal task performance.

4.4.3 *Discussion*

The analysis above shows that, within the constraints of our model, the experiment design is effective to reveal similar phenomena as the argued effects. Corresponding discussions can be outlined as follows:

(1) We break the traditional criteria of visual or tactile modality corresponded to the specific human senses, and reclassify the modalities into advantaged or disadvantaged modality without concerning specific sensory modalities. This new criteria reflect the individual difference of human senses and it is in accordance with the view of multisensory integration.

(2) The cross-modal difference is concerned an important influencing factor on the bimodal task performance, and the bimodal task performance here is compared to the advantaged unimodal task performance to be evaluated.

(3) The naturally clustered lower difference group shows a more observable performance enhancement effect than the higher difference group or the general average performance.

(4) From the grouped view, the average performance shows the phenomena described below:

Observed performance depression effect. For a group of subjects, if an individual unimodal signal can support much better task performance than the other modality, reduced bimodal task performance than the advanced-unimodal situation can be observed when synchronously presenting both modality signals to the group.

Observed performance enhancement effect. For a group of subjects, If an individual unimodal signal can support a similar task performance as the other modality, improved bimodal task performance than the advanced-unimodal situation can be observed when syschronously presenting both modality signals to the group.

5 Conclusion and future work

This paper introduces the newest view and related findings in cognitive science with the perspective of the multimodal human-computer interaction. Through combining the normalization model of multisensory integration and the Bayesian integration computation of sensorimotor task, we establish a multimodal interaction model based on neural computation. The model can describe user processing of multimodal signals in sensorimotor tasks and guide our exploration on the influence rules of multimodal presentation upon the task performance. The experimental result show the observed effects similar as the argued

effects. This shows a meaningful exploration of applied research study based on new basic scientific research findings.

The experimental result and discussion reveal new criteria of modality classification and the bimodal task performance evaluation. And the data analysis reflects the contribution of tactile feedback signal to the multimodal interactive task performance under this new criteria and evaluation. This demonstrates the value of this study for guiding the touch-included multimodal interaction design.

The future work mainly includes the following aspects:

(1) Further experiment: Considering the newest and mainstream technologies of touch feedback in further experiment, we want to find the cross modal influences between touch, vision and audition to improve multimodal interaction performance.

(2) Consider more cognitive findings: More discoveries will be derived by considering more cognitive findings related to multimodal interaction. Several existing laws, phenomena, derivations, and models, with the scope of cross-modal influence, may contribute to different research purposes.

(3) Build modeling framework: The methodology of modeling multimodal interaction needs to be investigated. A refined operable modeling architecture will be built such that different levels of canonical neural computations can be properly integrated to improve multimodal interaction design.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61232013, 61422212, 61303162) and National High Technology Research and Development Program of China (Grant Nos. 2015AA020506, 2015AA016305).

Conflict of interest The authors declare that they have no conflict of interest.

References

1 Jaimes A, Sebe N. Multimodal human-computer interaction: a survey. Comput Vis Image Und, 2007, 108: 116–134
2 Li M Z, Dai G Z, Dong S H. Software model and interaction algorithm of multimodal interface (in Chinese). Chinese J Comput, 1998, 21: 111–118
3 Salvendy G. Handbook of Human Factors and Ergonomics. Hoboken: John Wiley & Sons, 2012
4 Ware C. Information Visualization: Perception for Design. San Francisco: Morgan Kaufmann Publishers Incorporated, 2013
5 Shneiderman B, Plaisant C, Cohen M, et al. Designing the User Interface: Strategies for Effective Human-Computer Interaction. 5th ed. Upper Saddle River: Addison-Wesley, 2009
6 Wang H A, Ma C X. Interactive multi-scale structures for summarizing video content. Sci China Inf Sci, 2013, 56: 052108
7 Gallace A, Spence C. In Touch With the Future: the Sense of Touch From Cognitive Neuroscience to Virtual Reality. Oxford: Oxford University Press, 2014
8 Gallace A, Tan Z H, Spence C. The body surface as a communication system: the state of the art after 50 years. Presence, 2007, 16: 655–676
9 Bau O, Poupyrev I, Israr A, et al. TeslaTouch: electrovibration for touch surfaces. In: Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology. New York: ACM Press, 2010. 283–292
10 McGrath B J, Estrada A, Braithwaite M B, et al. Tactile Situation Awareness System Flight Demonstration. Army Aeromedical Resaerch Lab Fort Rucker AL Technical Report USAARL 2004-10. 2004
11 Lu L, Tian F, Dai G Z, et al. A study of the multimodal congnition and interaction based on touch, audition and vision (in Chinese). J Comput Aided Design Comput Graph, 2014, 26: 654–661
12 McGurk H, MacDonald J. Hearing lips and seeing voices. Nature, 1976, 264, 746–748
13 Mayer R E. Multimedia learning. Psychol Learn Motiv, 2002, 41: 85–139
14 Kieras D E, Meyer D E. An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. Human-Comput Interact, 1997, 12: 391–438
15 Ohshiro T, Angelaki D E, de Angelis G C. A normalization model of multisensory integration. Nature Neurosci, 2011, 14: 775–782
16 Carandini M. From circuits to behavior: a bridge too far? Nature Neurosci, 2012, 15: 507–509
17 Vroomen J, Keetels M. Perception of intersensory synchrony: a tutorial review. Atte Perce Psycho, 2010, 72: 871–884
18 Calvert G A, Spence C, Stein B E. The Handbook of Multisensory Processes. Cambridge: MIT Press, 2004
19 Ernst M O, Bülthoff H H. Merging the senses into a robust percept. Trends Cogn Sci, 2004, 8: 162–169
20 Stein B E, Stanford T R, Rowland B A. The neural basis of multisensory integration in the midbrain: its organization and maturation. Hearing Res, 2009, 258: 4–15
21 Carandini M, Heeger D J. Normalization as a canonical neural computation. Nature Rev Neurosci, 2012, 13: 51–62

22 Stein B E, Stanford T R. Multisensory integration: current issues from the perspective of the single neuron. Nature Rev Neurosci, 2008, 9: 255–266
23 Molholm S, Ritter W, Murray M M, et al. Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. Cogn Brain Res, 2002, 14: 115–128
24 Baranek G T. Efficacy of sensory and motor interventions for children with autism. J Autism Develop Disord, 2002, 32: 397–422
25 Brown S, Shankar R, Smith K. Borderline personality disorder and sensory processing impairment. Progress Neurol Psychiat, 2009, 13: 10–16
26 Stein B E, Meredith M A. The Merging of the Senses. Cambridge: MIT Press, 1993
27 Meredith M A, Stein B E. Spatial determinants of multisensory integration in cat superior colliculus neurons. J Neurophysiol, 1996, 75: 1843–1857
28 Meredith M A, Nemitz J W, Stein B E. Determinants of multisensory integration in superior colliculus neurons. I. temporal factors. J Neurosci, 1987, 7: 3215–3229
29 Meredith M A, Stein B E. Interactions among converging sensory inputs in the superior colliculus. Science, 1983, 221: 389–391
30 Spence C, Driver J. Crossmodal Space and Crossmodal Attention. Oxford: Oxford University Press, 2004
31 Gillmeister H, Eimer M. Tactile enhancement of auditory detection and perceived loudness. Brain Res, 2007, 1160: 58–68
32 Heeger D J. Normalization of cell responses in cat striate cortex. Visual Neurosci, 1992, 9: 181–197
33 Körding K P, Wolpert D M. Bayesian integration in sensorimotor learning. Nature, 2004, 427: 244–247
34 Ernst M O, Banks M S. Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 2002, 415: 429–433
35 Sternberg R. Cognitive Psychology. 3rd ed. Carolina: Wadsworth Publishing, 2003