IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 February 2024, accepted 7 March 2024, date of publication 11 March 2024, date of current version 15 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3375917

RESEARCH ARTICLE

Responsive-ExtendedHand: Adaptive Visuo-Haptic Feedback Recognizing Object Property With RGB-D Camera for Projected Extended Hand

YUSHI SATO, (Graduate Student Member, IEEE), DAISUKE IWAI, (Member, IEEE), AND KOSUKE SATO, (Member, IEEE)
Graduate School of Engineering Science, Osaka University, Toyonaka 560-8531, Japan

Corresponding author: Yushi Sato (y.sato@sens.sys.es.osaka-u.ac.jp)

This work was supported in part by the Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) Research Fellows under Grant JP23KJ1454; and in part by Japan Science and Technology Agency (JST), the Establishment of University Fellowships Toward the Creation of Science Technology Innovation, Japan, under Grant JPMJFS2125.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Research Ethics Committee of Osaka University under Application No. R2-28.

ABSTRACT The ExtendedHand interface displays computer graphics (CG) hand images in real space from a projector, allowing the user to visually point at and touch real objects that are out of their physical reach. Furthermore, when the projected CG hand (extended hand) touches an object, the user can feel the tactile sensation of the object through pseudo-haptics by giving the extended hand visual effects that emphasize the action. In the previous psychological study, the human operator had to manually assign the location and shape of objects and the intensities of their visual effects in advance in order to emphasize the appropriate visual effect for the object touched by the extended hand. To increase practical feasibility, we propose an adaptive system that utilizes an RGB-D camera and deep neural networks to generate the appropriate visual effects automatically and apply them to the projected extended hand. By employing U-Net to generate the appropriate intensities of the visual effects from the captured color and depth images, the system can estimate the appropriate visual effects for objects without pre-setting them. The user evaluation results showed that the proposed system allowed users to naturally perceive the tactile sensation of objects at a rate of 44%, instead of the manual rate of 49%.

INDEX TERMS Augmented reality, body augmentation, deep neural network, pseudo-haptics.

I. INTRODUCTION

Various initiatives are underway to utilize technology to enhance human physical and perceptual capabilities, enabling individuals to accomplish tasks and possess abilities that were previously unattainable [1], [2]. One such initiative is ExtendedHand, which visually extends the user's hand in everyday situations [3], [4]. This interface amplifies and reflects the user's hand movements in the movements of a computer graphics (CG) hand and projects them into real space using a projector. As a result, users can intuitively point to objects that are out of reach through the projected CG hand (referred to as the projected extended hand). Several applications of ExtendedHand have been proposed, such as facilitating communication between people and interacting with appliances by employing Internet of Things [4]. However, the user only receives visual information that the extended hand is projected onto objects (referred to as the projected extended hand

The associate editor coordinating the review of this manuscript and approving it for publication was Zeev Zalevsky.

touching objects). If the user could also perceive the tactile information of the objects, they would be able to experience previously impossible things, such as touching objects that are typically inaccessible, like museum exhibits.

One method of addressing the difference in tactile information between the projected extended hand and the actual hand is to use a haptic feedback device. This device can provide the same tactile stimulation as touching an object with actual hands [5], [6]. However, this approach requires preparing and wearing a dedicated haptic device, which limits the situations in which it can be used. As an alternative, Sato et al. [7] proposed a method that does not require haptic feedback devices. They introduced a technique for generating pseudo-haptics [8] by incorporating visual effects, such as vibrating the fingertips of the extended hand when it comes into contact with an object. They found that these visual effects can be used to perceive an object's unevenness, smoothness, and softness. However, their research was conducted as a psychological experiment to induce pseudo-tactile sensations; the position and properties of objects were known, and the application to practical situations where objects with various properties exist in various locations was not considered.

In this paper, we introduce a new function that senses the usage scene, recognizes information about the location and type of objects online, and adaptively applies the appropriate visual effect to the object touched by the projected extended hand. This enables the user to naturally perceive the tactile sensation of the touched object, even without prior information about the objects in the scene. We call the proposed system Responsive-ExtendedHand, which enhances the real-world applicability of ExtendedHand. To realize this system, we use an RGB-D camera to observe objects' shape and surface texture near the extended hand. We then employ U-Net [9] to estimate appropriate visual effects online based on the RGB-D images obtained. In this paper, we present the construction of Responsive-ExtendedHand and clarify its performance through a user study.

II. RELATED WORK

A. TACTILE FEEDBACK OF UNTOUCHABLE OBJECTS

Several studies have been conducted to make people feel as if they are touching objects that they cannot touch with their physical hands, such as remote or distant objects, by combining a substitute hand, such as CG hands or robotic hands, with haptic feedback devices [10], [11]. In the context of ExtendedHand, which is the focus of this study, Tanabe et al. [5] and Watanabe et al. [6] provided tactile stimuli to the user's hand using a haptic feedback device when the projected extended hand comes into contact with objects, thereby making the user perceive a sense of touch. Furthermore, Matsui et al. [12] and Sato et al. [7] have applied pseudo-haptic feedback [8], which generates haptic information from visual information, to ExtendedHand and have proposed a method to present tactile sensations of objects without the need for haptic feedback devices. These studies applied visual effects such as vibrating the fingertips or increasing the movement speed of the projected extended hand when it touched an object. This created the perception of tactile sensations, such as unevenness or smoothness, for the user. By providing tactile sensations of objects in ExtendedHand, users can not only perceive the characteristics of objects that are physically out of reach but also enhance their sense of ownership towards the projected extended hands [6].

In order to provide appropriate visual effects and haptic feedback based on the objects that the virtual hand interacts with, the system needs to have prior information about the positions, types, and characteristics of objects in the scene. In virtual reality (VR) spaces, this information is pre-modeled and stored as a scene model. However, the ExtendedHand interface running in mixed reality (MR) spaces requires online recognition and acquisition of object information at different locations in the real environment. Previous studies on ExtendedHand mainly focused on the user's psychological aspects, assuming that both information about object positions and suitable feedback are already known. However, these studies did not consider its applicability in practical situations where objects are present in different locations within an MR scene.

B. SCENE RECOGNITION USING DEEP LEARNING

When recognizing scene information, it is common to create an observation system using a sensor such as an RGB camera. The sensor values obtained are then used to extract and estimate the desired target information. Deep learning methods have gained significant attention in recent years for these purposes. Various approaches have been proposed to utilize deep learning to estimate object categories in RGB images. These approaches include methods that predict a single category for the entire image [13], methods that estimate categories for multiple objects in the image [14], and methods that estimate categories for each pixel in the image [9]. Furthermore, diverse estimation methods have been developed for specific categories. For example, some methods predict a universal set of 1,000 categories [13], while others focus on narrower domains, such as estimating 23 types of materials [15]. This diversity allows for a wide range of estimation possibilities, depending on the system's specific needs, as long as large-scale training datasets are available.

For ExtendedHand, it may be possible to estimate appropriate feedback based on the object touched by the projected extended hand using a deep learning framework. In particular, for tactile stimulus feedback [5], [6], vibration data from tracing an object can be used as appropriate tactile stimulus feedback based on the findings of previous studies [16], [17]. Several studies have already published large datasets of objects and vibration data when tracing them [18], [19]. However, in the case of visual effect feedback [7], [12], there is currently no dataset available that combines objects and visual effects. Additionally, research findings and the data collection experiment described in Section IV-B indicate that suitable visual effects for the same object highly rely on user preferences. Thus, creating a large dataset with multiple users

and training the network on that dataset does not guarantee high accuracy.

In this paper, we present a personal user system that aims to estimate appropriate visual effects for an object and apply them to the projected extended hand when it touches the object. To achieve this, we utilize RGB-D images and train a network on customized datasets consisting of object and visual effect data for each individual. While we rely on established deep learning techniques and a dataset of approximately 100 images per individual, we realize the system that can make users naturally perceive the tactile sensation of the object touched by the projected hand without prior information about various objects.

III. RESPONSIVE-EXTENDEDHAND

A. SYSTEM DESIGN

We present an overview and system flow of Responsive-ExtendedHand in Fig. 1. When the user moves their hand on a touch panel, the movement is amplified and reflected in the motion of the extended hand. The extended hand is projected onto a real scene using a video projector. An RGB-D camera captures the area surrounding the projected extended hand. When the system detects that the projected extended hand is overlapping an object in the RGB-D image, it adds visual effects suitable for the object to the projected extended hand and its surrounding area. The user can experience the tactile sensation of the object by seeing the projected extended hand with the visual effects, even though their hand is touching the touch panel [7].

Although the appropriate visual effects for object characteristics vary depending on user preferences, the proposed system fundamentally focuses on the following four situations based on previous studies [7], [12], as illustrated in Fig. 2:

(a) Bending-finger effect for an object's height difference,
(b) Shaking-finger effect for an uneven object,
(c) Increasing-speed effect for a slippery object,
(d) Deforming-object effect for a soft object.

B. SYSTEM FLOW

Responsive-ExtendedHand consists of two components: (A) Reflecting the user's hand movement and gestures onto the projected extended hand (green color area of the process flow in Fig. 1); and (B) Adding the appropriate visual effects to the extended hand by analyzing the scene (pink color area in Fig. 1). For component (A), we utilize ExtendedHand [4], which measures the user's hand movement from a touch panel input. Component (B) is further divided into the following four processes:

(B)-1 Visual sensing of the scene area around the projected extended hand,
(B)-2 Extraction of objects' physical properties from the sensor values,
(B)-3 Estimation of the appropriate visual effect based on the object's physical properties,
(B)-4 Modulation of the virtual hand image according to the estimated visual effect.

Here, processes (B)−2 and (B)−3 can be combined into a single process using a deep learning approach, if data on the relationship between the sensor values and the appropriate visual effect are available. These processes are explained in detail in the following.

1) AREA SENSING

We use an RGB-D camera as a sensor to capture the scene, which can measure the area around the projected extended hand without physical contact. This camera can extract material information from RGB color images. Additionally, it can gather information about objects' shapes and surface structures unaffected by texture or shading from Depth images. These features are essential for distinguishing object regions and determining the appropriate visual effects.

It is important to note that solely relying on RGB-D images makes it impossible to differentiate objects with similar appearances and shapes but varying hardness. The system prioritizes making users feel they are naturally touching objects rather than conveying the proper physical properties. Therefore, the system configuration solely depends on an RGB-D camera, which plays a role similar to the user's eyes.

The system clips only the projection area after geometrically transforming the captured RGB-D image using a pre-prepared pixel-to-pixel correspondence matrix between the RGB-D camera and the projector. This study limits the target object to a thin planar object and employs a homography transformation matrix as the correspondence matrix.

2) VISUAL EFFECT MAP GENERATION

The proposed system utilizes a deep learning framework to generate visual effect maps from the clipped RGB-D image. These maps determine the intensities of the visual effects for each pixel of the clipped RGB-D image (see Fig. 1). In this system, we utilize U-Net [9] to generate the visual effect maps (referred to as the *visual effect generation networks*). U-Net is a neural network that performs pixel-by-pixel segmentation of image input. Notable features of U-Net include its skip-connection structure, which accurately preserves boundary information for objects in the image. Additionally, U-Net can achieve high precision in identification even with limited data by utilizing data augmentation [9]. Considering that these features align with the requirements of the proposed system, we have chosen U-Net. This system uses separate networks for each visual effect to ensure easy scalability for potential additional types of visual effects in the future. In this system, the encoder and decoder layers of U-Net consist of eight layers each. The output layer uses a Sigmoid function to output values in the range of $[0, +1]$.

First, we resize the clipped RGB-D image to 256×256 pixels and then normalize the pixel values to the range of $[-1, +1]$. This normalized image is then used as the input for each network. Each network generates a visual effect map that holds the intensity values $[0, +1]$ of the corresponding visual

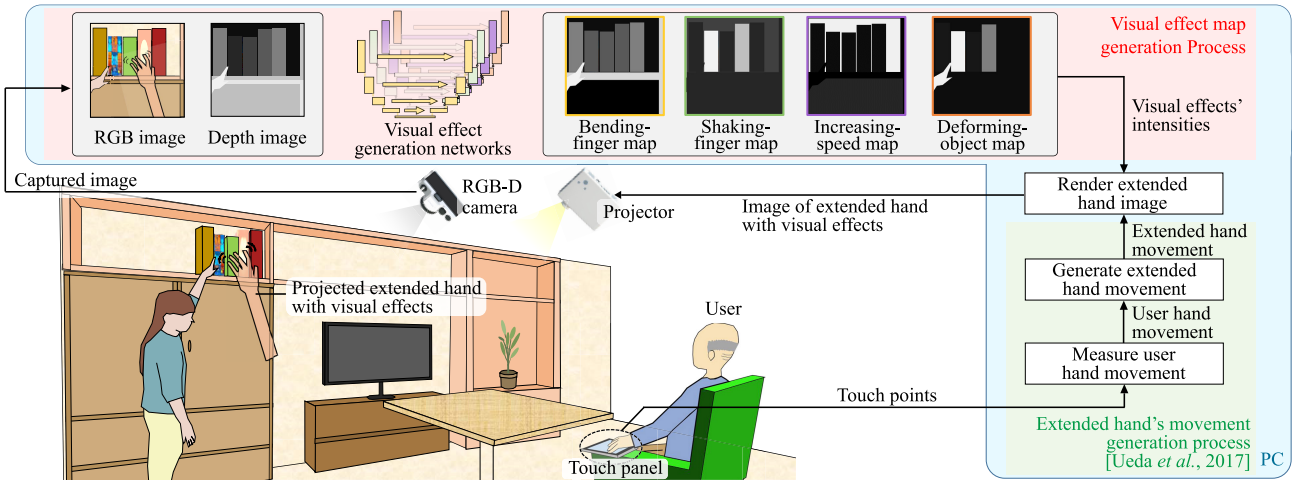


FIGURE 1. Overview and process flow of Responsive-ExtendedHand. The system generates visual effects suitable for the object being touched by the user-operated projected extended hand by employing an RGB-D camera and deep learning framework. This enables the user to feel the tactile sensation of the object through pseudo-haptics by viewing the projected extended hand with the visual effects, even without prior object information in the scene.

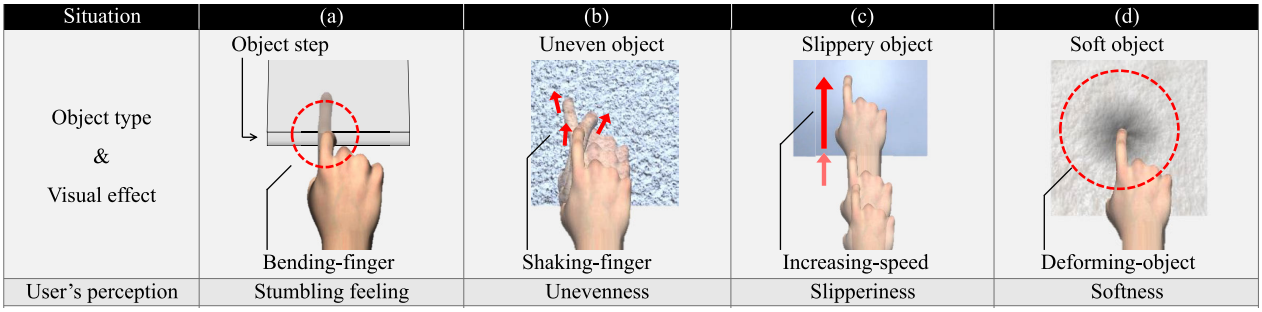


FIGURE 2. Visual effects that are applied when the projected extended hand touches an object. (a) Bending-finger effect for an object's height difference [12], [20], (b) Shaking-finger effect for an uneven object [7], (c) Increasing-speed effect for a slippery object [7], and (d) Deforming-object effect for a soft object [7].

effect for each pixel. The methodology for collecting training data and the training process is explained in Section III-C.

3) VISUAL EFFECT ADDITION

To apply visual effects to the projected extended hand, the system retrieves the pixel value from each visual effect map that corresponds to the fingertip position of the projected extended hand. The system then applies the corresponding visual effect with an intensity that matches the pixel value to the extended hand. If there are multiple types of visual effects with non-zero intensity values, the proposed system combines them. The Bending-finger effect is specifically designed to be applied only at object boundaries. This is accomplished by applying the effect only when the pixel value corresponding to the fingertip position of the extended hand changes by more than a threshold value (set empirically to 0.1) compared to its value in the previous frame.

C. TRAINING OF VISUAL EFFECT GENERATION NETWORKS

As mentioned in Section III-B2, training the *visual effect generation networks* requires a dataset of RGB-D images and their corresponding visual effect maps. The four visual effects shown in Fig. 2 are exaggerated representations of the physical phenomena that occur when an object is touched by a physical hand, which differ from the actual physical phenomena. Furthermore, the dataset collection experiment described in Section IV-B shows that the appropriate visual effect for the same object varies depending on the user's preference. Therefore, in this study, the system is configured for each user, and a dataset is prepared for each individual user.

A user follows a specific process to create the dataset, as illustrated in Fig. 3(a). They creates visual effect maps based on their preference for each object on the projection surface. This involves defining the object regions in the RGB-D images and setting appropriate visual effects intensities. The user replaces the objects on the projection surface with different types of objects for a limited number of iterations to complete the dataset. Subsequently, the dataset is expanded through the use of data augmentation [21]. During network training, the RGB-D images are inputs, while the corresponding visual effect maps serve as the ground truth (Fig. 3(b)).

IV. SYSTEM IMPLEMENTATION

We implemented the prototype system of Responsive-ExtendedHand based on Section III.

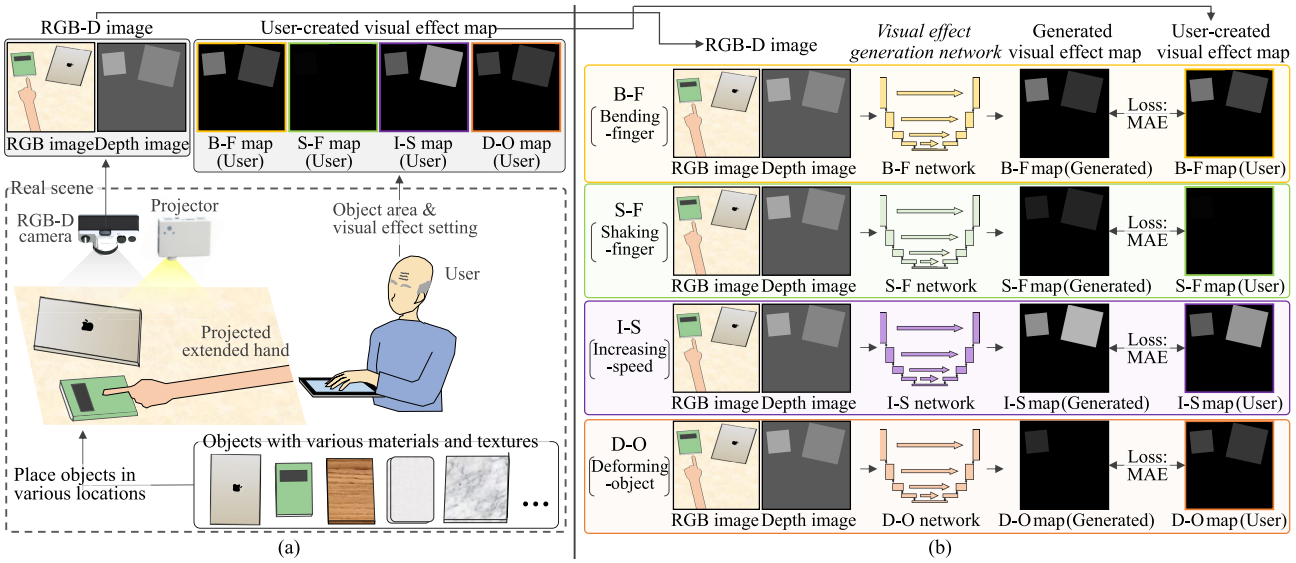


FIGURE 3. Procedure for training *visual effect generation networks*. (a) Creation of the training dataset. The user places different objects in the scene and configures the object area and appropriate visual effects for each object. The system stores the paired data of the captured RGB-D image and the user-created visual effect maps. (b) Training of the *visual effect generation networks*. The system trains each network using the RGB-D images as input and the user-created visual effect maps as ground truth.

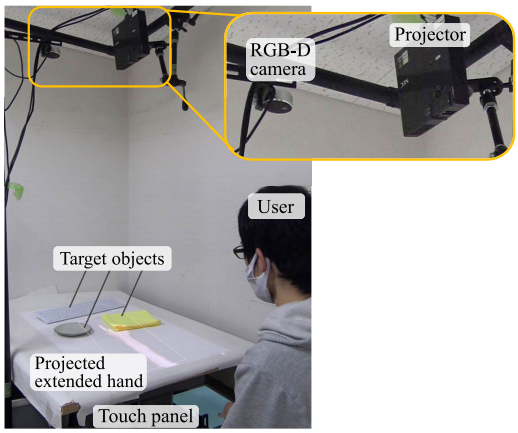


FIGURE 4. Appearance of the implemented system. The extended hand is projected onto a white table from a projector mounted on the ceiling. An RGB-D camera mounted next to the projector captures an RGB-D image of the projection area.

A. HARDWARE CONFIGURATION

Fig. 4 shows the appearance of the implemented system. The user used the projected extended hand on a white tabletop in this system. A projector (NEC, NP-L51WJD) was mounted on the ceiling and projected images onto a 540 mm × 910 mm area on the tabletop at 60 fps. An RGB-D camera (Intel, RealSense L515) next to the projector captured the projection surface at 30 fps. A touch panel (Microsoft, Surface Pro 3) placed beneath the tabletop enabled the user to manipulate the projected extended hand. The C/D (control-display) ratio was empirically set at 1:5. In other words, when a user moved their hand by 10 mm on the touch panel, the projected extended hand on the tabletop would move by 50 mm. Another PC (Microsoft, Surface Book 3) was employed to generate visual effect maps, control the extended hand's movements, and render the projection images.

B. CREATION OF TRAINING DATASET

In this implementation, 15 participants, aged 21 to 24, created datasets for training the *visual effect generation networks*. Each participant created 105 data points.

The experiments conducted in this section and Section V were approved by the Research Ethics Committee of Osaka University (No. R2-28). Additionally, we obtained written informed consent from each participant.

1) VISUAL EFFECT

We linearly normalized the intensity (degree of change) for each of the four visual effects shown in Fig. 2 within the range of [0, +1]. We refer to these intensities as t_{B-F}, t_{S-F}, t_{I-S}, and t_{D-O}, respectively. At the minimum intensity ($t = 0$), the corresponding visual effect was not applied. On the other hand, at the maximum intensity ($t = 1$), the corresponding visual effect change was overemphasized. In this case, almost all participants perceived the change in the projected extended hand as being caused by factors other than the characteristics of the touched object. The specific changes produced at minimum and maximum intensity were determined in Table 1 using the design parameters format from previous studies [7], [20].

2) TARGET OBJECT

Based on relevant research [15], [22], we selected seven commonly used indoor materials: ceramic, fabric, metal, paper, plastic, stone, and wood. For each material, we chose five objects with distinct surface textures. As a result, the 35 objects shown in Fig. 5 were prepared as objects that the projected extended hand touched. In this study, we excluded objects with low reflectance or significant height variations that cannot be effectively corrected using homography

TABLE 1. Design parameter values for the visual effects [7], [20] at mavimum and minimum intensity.

Bending -finger [20]	l_{th} [mm] (Amount of finger joint bending)			
$t_{B-F} = 0$	0.0			
$t_{B-F} = 1$	100.0			
Shaking -finger [7]	A_{real} [mm] (Finger tip amplitude)	λ [mm] (Finger vibration period)		
$t_{S-F} = 0$	0.0	10.0		
$t_{S-F} = 1$	2.0	10.0		
Increasing -speed [7]	γ (Rate of speed increase)			
$t_{I-S} = 0$	1.0			
$t_{I-S} = 1$	2.0			
Deforming -object [7]	r [mm] (Area)	$time$ [ms] (Duration)	d [mm] (Depth)	d_{shade} (Shade darkness)
$t_{D-O} = 0$	0.0	400	0.1	0.3
$t_{D-O} = 1$	100.0	200	1.7	0.3

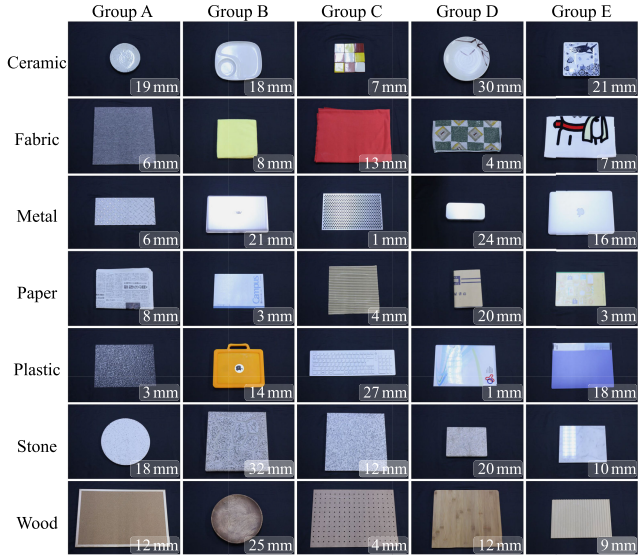


FIGURE 5. 35 different objects used in training and evaluation. The size of each image is approximately 500 mm in width and 300 mm in height. The numerical values indicate the maximum thickness of the objects.

transformation. The white tabletop was also considered the background and not included as part of the target objects.

3) COLLECTION PROCEDURE

Participants were given the task of adjusting the appropriate intensities of visual effects for objects. To perform this task, participants used their index finger to operate the projected extended hand at a speed of approximately 200 mm/s. Ample practice was provided beforehand to ensure participants could achieve this speed.

At the beginning of each trial, an experimenter placed two or three objects on the white tabletop. These objects belonged to the same group, as indicated in Fig. 5, and their placement locations were randomly determined by the system to avoid overlap. The system then instructed the participant to trace one of the objects using the projected extended hand. As the projected extended hand overlapped with the object,

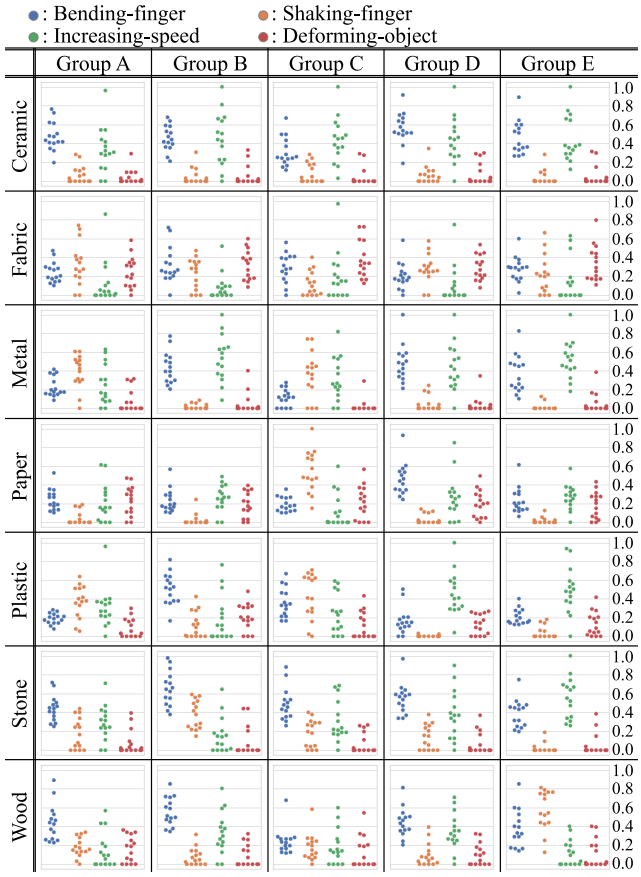


FIGURE 6. Distribution of the intensities of the visual effects set by the participants for each of the 35 objects. Each dot represents an individual participant. The median values are used since each participant sets visual effects three times for each object.

four visual effects were added. The participant adjusted the intensity of each of the four visual effects by operating the position of the four sliders on the MIDI controller (Worlde, EasyControl.9). The goal was for the participant to set the four intensities at which they felt most natural touching the object with the projected extended hand.

Once the participant decided on the visual effects, the system recorded the RGB-D image and the intensities of the set visual effects. After the recording, the participant was instructed to perform the same task on the remaining objects on the table. This process continued until the task was completed for all the objects. Then, a new set of objects was placed for the next round of tasks.

Each participant performed this task three times for each of the 35 objects, resulting in a total of 105 trials. The entire task, including explanation time and breaks, took approximately two hours to complete. The order in which the objects were touched and the combinations of objects placed on the table were randomized.

4) CREATED DATASET

We collected 105 RGB-D images and their corresponding visual effect maps per participant. Fig. 6 shows the distribution of the intensities of the four visual effects that each

participant set for each object. Since each participant set the intensities three times for each object, we used the median value as a representative measure. These results highlight significant variations in the intensities set by each participant for the same object, especially for the Increasing-speed effect.

C. TRAINIG OF VISUAL EFFECT GENERATION NETWORKS

We trained the *visual effect generation networks* using the dataset created in Section IV-B. As mentioned in Section III-C, for this study, we trained separate networks tailored to each participant using datasets created by each participant.

Considering the practical application scenarios of the proposed system, it is not feasible to require users to pre-set appropriate visual effects for all objects in the scene. Therefore, the system needs to accommodate two categories of objects: **known objects**, which were included in the data for network training, and **unknown objects**, which were not included in the network training. To evaluate both known and unknown objects in the user study in Section V, we used data from 28 out of 35 objects for network training. The remaining seven objects were kept unknown for the purpose of evaluation.

1) TRAINING CONDITION

For each participant's 105 data points, we utilized 84 data from three evaluations of 28 out of 35 objects for training. We selected these 28 objects from four out of the five groups shown in Fig. 5. Therefore, our training data consisted of four instances of each of the seven materials. The selection of the four groups was balanced across participants and randomized.

We expanded the dataset from 84 to 2,520 data points, increasing it thirty-fold using data augmentation techniques [21], such as brightness modulation and geometric transformations. Next, we trained each of the four *visual effect generation networks* using the expanded dataset. We used a batch size of 10 and employed the Adam optimization algorithm with a learning rate of 10^{-3}. We used the Mean Absolute Error (MAE) loss function and ran the training for 50 epochs. During each epoch, we used 20% of the training data as validation data.

2) PREDICTION RESULTS FOR UNKNOWN OBJECTS

We generated visual effect maps from RGB-D images of 21 data points (seven objects, each evaluated three times) excluded from the training using the trained networks for each participant. Fig. 7 illustrates examples of the generated visual effect maps. We computed the Mean Absolute Error (MAE) between the generated maps and the ground truth maps created by the participants. Additionally, we separated the MAE calculations into the the background area (where the white table appears in the RGB-D images) and the target object area (where the target objects appear in the RGB-D images). Table 2 presents these results. Furthermore, we computed the MAE for each of the 35 objects. Fig. 8 presents the results.

TABLE 2. MAEs were calculated for the entire map, as well as for the regions corresponding to the background and target object areas of the input RGB-D images, respectively. The values represent the mean and standard deviation.

Type of visual effects	MAE for the whole map	MAE for the background area of the map	MAE for the target object area of the map
Bending-finger map	0.05 ± 0.03	0.01 ± 0.01	0.17 ± 0.09
Shaking-finger map	0.06 ± 0.04	0.01 ± 0.01	0.19 ± 0.13
Increasing-speed map	0.07 ± 0.03	0.01 ± 0.01	0.21 ± 0.11
Deforming-object map	0.04 ± 0.03	0.01 ± 0.01	0.12 ± 0.09

The average MAE for the background area across all four visual effects generation networks was 0.01. This suggests that the networks were capable of recognizing the background region (the white tabletop). On the other hand, the average MAE for the target object area ranged between 0.12 and 0.21 across the four networks. For the target object area, the networks must not only identify the object's presence but also recognize its characteristics and determine the appropriate intensities of the visual effects. Therefore, it is inevitable that the MAE for the target object area was worse than that of the background area.

Focusing on individual objects, the average MAE values for most objects ranged from 0.1 to 0.3 across the four networks, shown in Fig. 8. Since there were no materials with notably large or small MAE values, it is suggested that the four networks do not exhibit a particular proficiency or deficiency for specific object material types.

However, the MAE for objects in Group C, particularly paper, metal, and plastic materials, was notably poorer than that of other objects in the Shaking-finger generation network. One potential explanation for this observation is that there were relatively many objects with uneven surfaces in Group C. In contrast, other groups had fewer objects with such uneven surfaces (such as stone materials in Groups A and B, metal materials in Group A, and wood materials in Group E). The MAE values might have been compromised because the Shaking-finger's intensity was estimated for uneven objects that were not extensively contained in the training data of the network.

Comparing the types of visual effects, the MAE values of the deforming object generation network were notably better than the others. This would occur because the intensity of the Deforming-object effect set by participants for each object was mostly 0.5 or below (see Fig. 6). As a result, the variance of the set Deforming-object's intensity for different objects was smaller than that of the other visual effects.

In this section, we have discussed the generation accuracy of the *visual effect generation networks* in terms of MAE values. However, how much these MAE values influence user perception is still unclear. This study aims to determine whether the proposed system can naturally convey the tactile sensation of objects to the user without prior object information. We will verify this aspect through the user study in Section V.

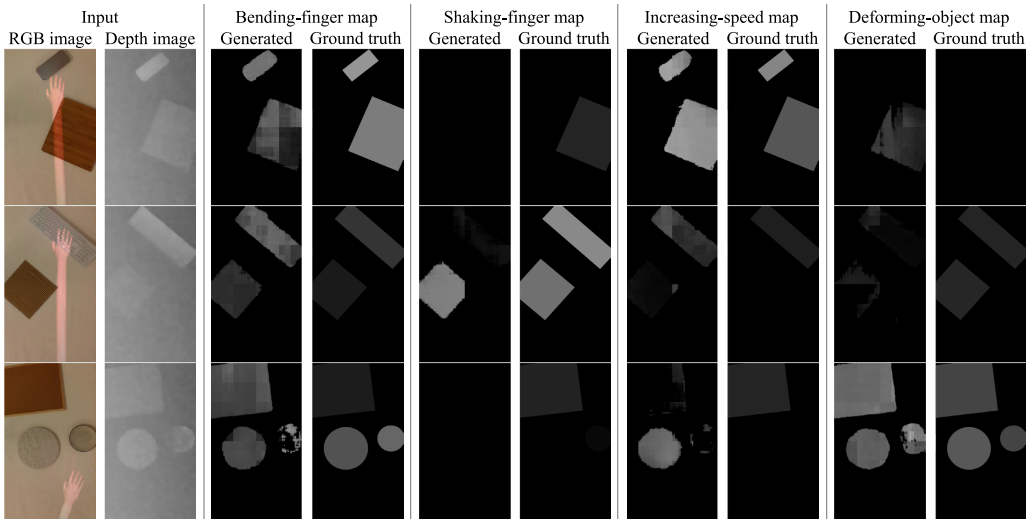


FIGURE 7. Examples of the generated visual effect maps. These maps were generated by the trained visual effect generation networks using RGB-D images that were not included in the training.

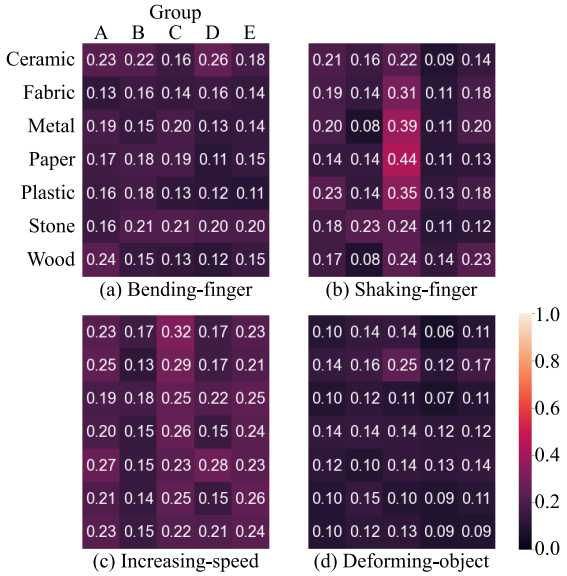


FIGURE 8. MAE results for each of the 35 objects. The values represent the mean.

D. ONLINE PROCESSING

We integrated the *visual effect generation networks*, trained in the previous section, into the prototype system shown in Fig. 4. Subsequently, we conducted evaluations in the environment depicted in Fig. 4. The time it took for the user's hand movement to be reflected in the motion of the projected extended hand was 150 ms. Shimada et al. [23] reported that users do not consciously notice delays below 200 ms, so the implemented system met this requirement.

In the implemented system (using a GPU: NVIDIA, GeForce GTX 1650), it took approximately 200 ms to generate a visual effect map with an image size of 256 × 256 pixels. The motion generation process for the projected extended hand and the visual effect generation process were handled in separate threads. Therefore, this delay did not affect the motion of the projected extended hand. This means that while providing visual effects to rapidly moving objects in the usage scene may be challenging, it is possible to provide suitable visual effects for relatively stationary objects with occasional changes in position or shape, even on less powerful PCs.

V. USER EVALUATION

We conducted a user study to assess the performance of the proposed system in a typical scenario where there is no prior information available about objects in the scene. This study aimed to determine whether users can naturally perceive the tactile sensations of objects touched by the projected extended hand.

A. CONDITION

1) PARTICIPANT

The participants in this experiment were the same 15 individuals who participated in the dataset creation described in Section IV-B.

2) VISUAL EFFECT ADDITION

We used the system implemented in Section IV-C to generate visual effects. Specifically, we trained the *visual effect generation networks* using data from 28 objects (four groups), as shown in Fig. 5. We will refer to this condition as the **Prop condition**.

Furthermore, for comparison, we introduced the following two conditions requiring the prior object information:

a: PERFECT CONDITION

In this condition, when the projected extended hand touched an object, the system provided the visual effects that were set by the respective participant for the object during the dataset creation in Section IV-B. We used the median value since each participant set the visual effects three times for each object.

b: CONST CONDITION

In this condition, when the projected extended hand touched an object, the system provided the same visual effects

regardless of the type of the touched object. The visual effects were the average values set by each participant for all objects during the dataset creation in Section IV-B.

3) TARGET OBJECT

As mentioned in section IV-C, we prepared two categories of objects to be touched by the projected extended hand: **Known objects**, which were included in the training data of the *visual effect generation networks*, and **Unknown objects**, which were not included.

Each category consisted of seven objects (corresponding to one group in Fig. 5), one for each of the seven materials. For known objects, one group was chosen from the four groups used during training. For unknown objects, one group that was not used during training was selected. The selection of each group was randomized to ensure balance among participants.

B. PROCEDURE

The experiment was conducted in the same environment described in Section IV-B, shown in Fig. 4. Initially, participants practiced manipulating the projected extended hand. Similar to Section IV-B, they used a single index finger to control the projected extended hand at a speed of approximately 200 mm/s. They received ample practice to become proficient in this operation. Following the practice session, participants repeated the following task:

Step 1: The experimenter arranged two or three objects on the white tabletop, ensuring that they did not overlap. The system randomly determined the types and placement of these objects.

Step 2: The system instructed the participant to touch one of the objects. The participant used the projected extended hand to touch and trace the indicated object. During this interaction, visual effects were applied to the projected extended hand under one of three conditions: Prop, Perfect, or Const. After the interaction, participants responded to the following two questions on a 7-point Likert scale (−3: Strongly disagree — +3: Strongly agree):

Q1: Did you feel as though you were touching the object naturally with the projected extended hand?

Q2: Did you perceive the tactile sensation of the object?

For Q1, participants were instructed to evaluate whether the appearance and movement of the projected hand overlapping the object were acceptable, rather than whether they resembled the appearance and movement of an actual hand touching the object. As mentioned at the beginning of this section, this study aimed to determine on whether participants could naturally perceive the tactile sensation of the object. We selected these questions because this criterion could be examined by analyzing the frequency of high scores for both Q1 and Q2.

Step 3: After answering the questions, participants were instructed to perform the same task on another object on the tabletop that they had yet to assess. When participants performed the task for all objects on the tabletop, they started from **Step 1** for another set of objects.

Each participant touched 14 objects (seven known and seven unknown) under each of the three visual effect addition conditions, resulting in a total of 42 times performing this task. The order of conditions was randomized and balanced across the participants. After completing all the tasks, participants verbally provided their impressions.

C. RESULTS

Fig. 9 presents the evaluation results for Q1 and Q2 in each condition. In this figure, the horizontal axis represents the scores for Q1 (−3 to +3), and the vertical axis represents the scores for Q2 (−3 to +3). Each cell shows the number of votes corresponding to the respective scores.

1) VISUAL EFFECT ADDITION FACTOR (Fig. 9(a))

This user study aimed to determine whether participants naturally perceived the tactile sensation of objects touched by the projected extended hand. Therefore, as described in Section V-B, we examined the rate of each participant who scored one or higher on both Q1 and Q2 in each condition (highlighted in the green box in Fig. 9(a)). The mean and standard deviation were as follows: Prop: 44.3%±22.8%, Perfect: 49.0%±23.7%, Const: 35.7%±24.2%. We performed an ANOVA with the visual effect addition as a factor. The ANOVA result showed a significant difference ($F(2, 14) = 3.51, p < 0.05$). Post-hoc multiple comparisons with Bonferroni correction revealed that the rate in the Perfect condition was significantly higher than in the Const condition ($p < 0.05$).

2) TARGET OBJECT FACTOR (Fig. 9(b))

We performed the same analysis of results for the target object factor, and the results were as follows: Known objects: 47.6%±30.1%, Unknown objects: 41.0%±21.4%. The t-test result did not reveal any significant differences ($t(14) = 0.97$, $p > 0.05$).

3) RESULTS FOR EACH OBJECT (Fig. 10)

We evaluated each of the 35 objects. We counted instances where both Q1 and Q2 received scores of 1 or higher. The results are shown in Fig. 10. Each object was evaluated twice by three participants under the Prop, Perfect, and Const conditions (For the Prop condition, three participants evaluated the objects once under the Known object condition and once under the Unknown object condition). Therefore, each object had a maximum of six assessments per condition.

4) PARTICIPANTS' COMMENTS

In the verbal feedback from the participants, all of them mentioned that the appearance of visual effects that matched the objects enhanced the sensation of touching them. However, in 12 cases, participants reported that the appearance of visual effects that did not match the objects felt unnatural (e.g., it was unnatural for the Deforming-object to appear when

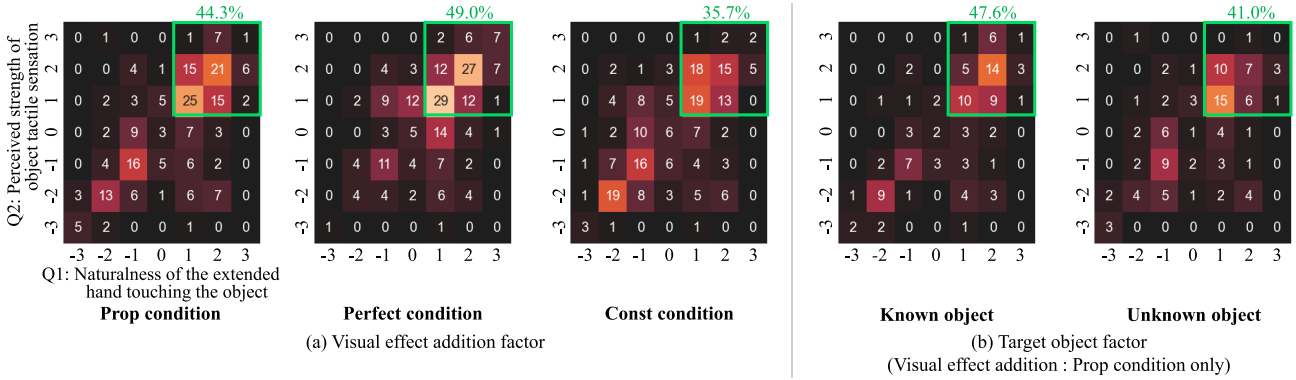


FIGURE 9. Results of participant evaluations. Each cell value represents the number of times the corresponding Q1 and Q2 were answered. The green percentages indicate the rate of participants who naturally perceived the tactile sensation of the objects (Q1>0 and Q2>0).

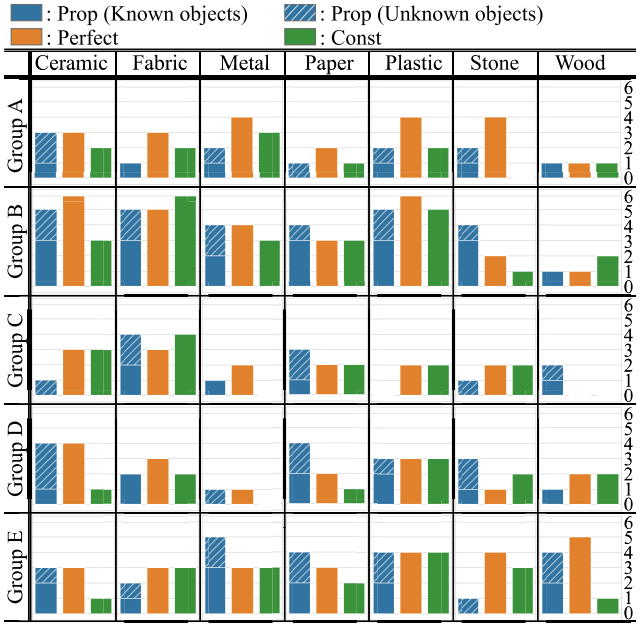


FIGURE 10. Results of participant evaluations for each of the 35 objects. The vertical axis on each graph represents the number of times participants naturally perceive the object' tactile sensation (Q1>0 and Q2>0). Each object was evaluated a total of six times under each condition, so the maximum value on the vertical axis is six.

touching a hard stone; or it was unnatural that the shaking finger did not appear for objects with uneven surfaces). Additionally, there were four reports indicating that the visual effect appeared in places where no object existed.

D. DISCUSSION

The proposed system aims to enable users to naturally perceive the tactile sensations of different objects touched by the projected extended hand without prior information about the objects. To assess this, we analyzed the rate of scores one or higher in both Q1 and Q2. The Perfect condition used the visual effects set by the participants for each object in Section IV-B. As a natural consequence, the Perfect condition had the highest average value of 49.0% among the three conditions. On the other hand, the average difference between the Prop and Perfect conditions was 4.7%, which was not statistically significant. This means that we cannot definitively conclude that there is no difference between the two conditions. It suggests that the proposed system (Prop condition) may perform worse than when object information is pre-set (Perfect condition).

However, the typical usage scenario for ExtendedHand does not provide information about the location and types of various objects in the scene. In these scenarios, the results showed that the proposed system could naturally make users perceive the tactile sensation of objects touched by the projected extended hand with high validity, with the preparation of about 100 data points. This is compared to the scenario where object information is provided in advance (Prop/Unknown Object condition: 41.0%, Perfect condition: 49.0%). Although the proposed system may be inferior to manually setting visual effects, it is considered the first example of generating pseudo-haptic sensations for unknown objects by incorporating online object recognition.

Examining the results for each object (Fig. 10), it is evident that several objects consistently obtained low scores regardless of the visual effect addition factor, such as the metal object in Group D and the wood object in Group A. This suggests that there is a limitation to the range of tactile sensations expressed by the four visual effects used in this study.

Although there are exceptions due to the small number of data, the results shown in Fig. 10 also indicate the following tendency: the Prop condition generally obtained slightly lower scores compared to the Perfect condition for all objects, rather than significantly lower for a specific material. This finding aligns with the results presented in Section IV-C2, where the MAE values ranged from 0.1 to 0.3 for all objects. In light of this, potential improvements could be achieved by refining the data augmentation techniques in the training data [24] or utilizing transfer learning approaches [21].

VI. CONCLUSION

In this paper, we proposed Responsive-ExtendedHand, which integrates scene observation using an RGB-D camera and online object recognition using deep learning techniques into ExtendedHand to adaptively estimate appropriate visual

effects for objects touched by the projected extended hand. The system aimed to allow the user to perceive the tactile sensations of the objects, even without prior information about the objects in the scene. The user evaluation results indicated that the proposed system performed slightly worse than the Perfect condition, which requires complete information about the location and type of the objects. However, it successfully enabled users to naturally perceive the tactile sensation satisfactorily without needing such information.

Future work will focus on generating appropriate visual effects for unspecified users by considering not only the RGB-D image but also the user's preferences. Additionally, this paper primarily addressed situations where few objects are sparsely distributed. However, we intend to expand our system's capabilities to handle situations where objects are densely distributed.

REFERENCES

[1] R. Raisamo, I. Rakkolainen, P. Majaranta, K. Salminen, J. Rantala, and A. Farooq, "Human augmentation: Past, present and future," *Int. J. Hum.-Comput. Stud.*, vol. 131, pp. 131–143, Nov. 2019.

[2] D. Prattichizzo, M. Pozzi, T. L. Baldi, M. Malvezzi, I. Hussain, S. Rossi, and G. Salvietti, "Human augmentation by wearable supernumerary robotic limbs: Review and perspectives," *Prog. Biomed. Eng.*, vol. 3, no. 4, Oct. 2021, Art. no. 042005.

[3] S. Ogawa, K. Okahara, D. Iwai, and K. Sato, "A reachable user interface by the graphically extended hand," in *Proc. 1st IEEE Global Conf. Consum. Electron.*, Oct. 2012, pp. 210–211.

[4] Y. Ueda, Y. Asai, R. Enomoto, K. Wang, D. Iwai, and K. Sato, "Body cyberization by spatial augmented reality for reaching unreachable world," in *Proc. 8th Augmented Human Int. Conf.*, Mar. 2017, pp. 1–9.

[5] N. Tanabe, Y. Sato, K. Morita, M. Inagaki, Y. Fujino, P. Punpongsanon, H. Matsukura, D. Iwai, and K. Sato, "FARFEEL: Providing haptic sensation of touched objects using visuo-haptic feedback," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces (VR)*, Mar. 2019, pp. 1355–1356.

[6] A. Watanabe, T. Uchida, D. Iwai, and K. Sato, "Hovering and contact representation of laser contour-based hand with swinging tablet PC for distant communication," in *Proc. 16th Int. Conf. Tangible, Embedded, Embodied Interact.*, Feb. 2022, pp. 1–7.

[7] Y. Sato, T. Hiraki, N. Tanabe, H. Matsukura, D. Iwai, and K. Sato, "Modifying texture perception with pseudo-haptic feedback for a projected virtual hand interface," *IEEE Access*, vol. 8, pp. 120473–120488, 2020.

[8] A. Lécuyer, S. Coquillart, A. Kheddar, P. Richard, and P. Coiffet, "Pseudo-haptic feedback: Can isometric input devices simulate force feedback?" in *Proc. IEEE Virtual Reality*, Mar. 2000, pp. 83–90.

[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)*, Oct. 2015, pp. 234–241.

[10] C. L. Fernando, M. Furukawa, T. Kurogi, S. Kamuro, K. Sato, K. Minamizawa, and S. Tachi, "Design of TELESAR V for transferring bodily consciousness in telexistence," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 5112–5118.

[11] T. Duan, P. Punpongsanon, D. Iwai, and K. Sato, "FlyingHand: Extending the range of haptic feedback on virtual hand using drone-based object recognition," in *Proc. SIGGRAPH Asia Tech. Briefs*, Dec. 2018, pp. 1–4.

[12] K. Matsui, K. Sato, and D. Iwai, "Pseudo haptic feedback to the operator by visual effect of the virtual hand," (Japanese), in *Proc. SICE SI*, Dec. 2018, pp. 377–380.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[15] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3479–3487.

[16] P. Wellman and R. D. Howe, "Towards realistic vibrotactile display in virtual environments," *Proc. ASME Dyn. Syst. Control Division*, vol. 57, no. 2, pp. 713–718, Nov. 1995.

[17] A. M. Okamura, J. T. Dennerlein, and R. D. Howe, "Vibration feedback models for virtual environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, vol. 1, May 1998, pp. 674–679.

[18] M. Strese, J.-Y. Lee, C. Schuwerk, Q. Han, H.-G. Kim, and E. Steinbach, "A haptic texture database for tool-mediated texture recognition and classification," in *Proc. IEEE Int. Symp. Haptic, Audio Vis. Environ. Games (HAVE)*, Oct. 2014, pp. 118–123.

[19] M. Strese and E. Steinbach, "Toward high-fidelity haptic interaction with virtual materials: A robotic material scanning, modelling, and display system," in *Proc. IEEE Haptics Symp. (HAPTICS)*, Mar. 2018, pp. 247–254.

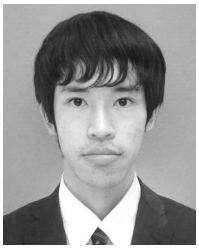
[20] Y. Sato, D. Iwai, and K. Sato, "A study on visual effects in projected extended hand interactions," (Japanese), in *Proc. 26th Annu. Conf. Virtual Reality Soc. Jpn.*, Sep. 2021, pp. 1–4.

[21] V. Suryamurthy, V. S. Raghavan, A. Laurenzi, N. G. Tsagarakis, and D. Kanoulas, "Terrain segmentation and roughness estimation using RGB data: Path planning application on the CENTAURO robot," in *Proc. IEEE-RAS 19th Int. Conf. Humanoid Robots (Humanoids)*, Oct. 2019, pp. 1–8.

[22] H.-S. Yeo, J. Lee, A. Bianchi, D. Harris-Birtill, and A. Quigley, "SpeCam: Sensing surface color and material with the front-facing camera of a mobile device," in *Proc. 19th Int. Conf. Human-Comput. Interact. Mobile Devices Services*, Sep. 2017, pp. 1–9.

[23] S. Shimada, Y. Qi, and K. Hiraki, "Detection of visual feedback delay in active and passive self-body movements," *Exp. Brain Res.*, vol. 201, no. 2, pp. 359–364, Mar. 2010.

[24] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1310–1319.



YUSHI SATO (Graduate Student Member, IEEE) received the B.S. and M.S. degrees from Osaka University, Japan, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree. His research interests include the areas of augmented reality and human–computer interaction. He is a JSPS Research Fellow DC2.

DAISUKE IWAI (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Osaka University, Japan, in 2003, 2005, and 2007, respectively. He was a Visiting Scientist with Bauhaus University Weimar, Germany, from 2007 to 2008, and a Visiting Associate Professor with ETH Zürich, Switzerland, in 2011. He is currently an Associate Professor with the Graduate School of Engineering Science, Osaka University. His research interests include spatial augmented reality and projector-camera systems.

KOSUKE SATO (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Osaka University, Japan, in 1983, 1985, and 1988, respectively. He was a Visiting Scientist with the Robotics Institute, Carnegie Mellon University, from 1988 to 1990. He is currently a Professor with the Graduate School of Engineering Science, Osaka University. His research interests include image sensing, virtual reality, and human interface.