Doctoral Dissertation

Effects of changes in food and environmental appearance by augmented/virtual reality on multisensory flavor perception

Kizashi Nakano

March 17, 2023

Graduate School of Science and Technology
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Science and Technology,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Kizashi Nakano

Thesis Committee:

Kiyoshi Kiyokawa

(Professor, Division of Information Science)

Hirokazu Kato

(Professor, Division of Information Science)

Hideaki Uchiyama

(Associate Professor, Division of Information Science)

Naoya Isoyama

(Assistant Professor, Division of Information Science)

Monica Perusquia-Hernandez

(Assistant Professor, Division of Information Science)

Effects of changes in food and environmental appearance by augmented/virtual reality on multisensory flavor perception [*]

Kizashi Nakano

Abstract

Gustation is a sensation resulting from a chemical reaction. Because of this, it is difficult to present artificially. In contrast, the taste we perceive when we eat food is perceived as flavor by integrating multiple senses, including sight and smell. For this reason, research is being conducted on the sensory presentation of gustation by changing the visual stimuli, which has some degree of established presentation. Previous studies have successfully presented the taste between specific food types using pre-created images. However, they could not cope with the complex deformation of food, and detailed verification of taste changes has yet to be performed. This dissertation aims to establish a visual change method to perceive different types of food and to investigate the effects of changes in visual information on multisensory flavor. We developed the following three types of gustatory manipulation methods: 1. Gustatory manipulation interface that changes the appearance of food into the appearance of different food (Chap. 3 and Chap. 4). 2. Application that superimposes only the eating region on the virtual environment to investigate the effect of changing the appearance of the surroundings to the virtual environment (Chap. 5). 3. Head-mounted display with an increased downward field of view to investigate the effect of visual information presented close to the mouth on the gustation (Chap. 6), and possibly to study the impact of changes in the avatar's body on the gustation (Chap. 7).

[*] Doctoral Dissertation, Graduate School of Science and Technology, Nara Institute of Science and Technology, March 17, 2023.

The results of the demonstration experiment revealed that gustatory manipulation by visual modulation can persistently change the taste and type of food and can present smell and food texture. Although changes in the appearance of the surrounding environment and the presentation of visual information close to the mouth did not affect gustation, we were able to construct the system necessary to verify the effect of visual modulation on gustation. These results show that we have succeeded in developing the techniques necessary for gustatory manipulation by visual modulation and making food perceived as different foods as intended.

Keywords:

Multisensory flavor perception, Multisensory integration, Gustatory manipulation, Cross-modal effect, Augmented reality, Virtual reality, Head-mounted display, Downward field of view, Generative adversarial network, Machine learning

Contents

List of Figures **viii**

1 Introduction **1**

1.1 A method of presenting the multisensory flavors of food 1

1.1.1 Change the appearance of food to different food intended . 3

1.2 Eating in the virtual world looks different from the real world . . 4

1.3 Alter the appearance of space close to our bodies 5

1.4 Contribution 7

2 Related Works **9**

2.1 Mechanism of human multisensory flavor perception 9

2.2 Gustatory manipulation by multisensory presentation 10

2.3 Food image generation using generative adversarial networks . . . 22

2.4 Eating while wearing HMD in the VE 25

2.5 Food Segmentation 29

2.6 HMDs with a wide-FoV 31

2.7 Information presented on the downward FoV 35

3 Manipulation of taste and type of food by changing its appearance **42**

3.1 GAN-based Real-time Food-to-Food Translation and Its Impact on Vision-induced Gustatory Manipulation 42

3.2 Effectiveness and Problems of 3D Model-based System 43

3.2.1 Food Selection 43

3.2.2 System Overview 44

3.2.3 Procedure 45

3.2.4 Discussion and Identified Problems 47

3.3 GAN-based Real-time Food-to-food Translation System 48
3.3.1 System Overview 49
3.3.2 Client Module 50
3.3.3 Server Module 51
Preprocessing 52
StarGAN 52
3.4 Experiment 54
3.5 Results and Discussion 56
3.5.1 Overview 56
3.5.2 Noodle Conditions 57
3.5.3 Rice Conditions 61
3.5.4 Overall Discussion 64
3.6 Limitations 66
3.6.1 Food Types and Participants 66
3.6.2 Experimental Protocol 67
3.6.3 Image Translation Quality 67
3.6.4 System Latency 68
3.6.5 Difficulty of Eating 69
3.7 Conclusion 69

4 Persistence of cross-modal effectse between vision and gustation 70

4.1 Gustatory Manipulation by GAN-based Visual Modulation Persists for Several Bites of Food 70
4.2 Experiment 72
4.2.1 Overview 72
4.2.2 Procedure 74
4.3 Results and Discussions 75
4.3.1 Overview 75
4.3.2 Noodle Conditions 76
Results 76
Discussion 79
4.3.3 Rice Conditions 85
Results 85
Discussion 89

4.3.4 Correlation between Noodle and Rice Conditions 92
4.3.5 Overall Discussion 93
4.4 Limitations 94
4.4.1 Food Types 94
4.4.2 Participants 95
4.4.3 Experimental Protocol 95
4.4.4 Image Translation Quality 95
4.4.5 System Latency 96
4.4.6 Difficulty of Eating 96
4.5 Conclusions 96

5 Effects on multisensory flavor perception when the surrounding environment is changed to a virtual environment 98
5.1 A Video See-through Food Overlay System for Eating Experience in the Metaverse 98
5.2 System 100
5.3 Experiment 102
5.3.1 Overview 102
5.3.2 Hypotheses 103
5.3.3 Procedure 104
5.3.4 Questionnaire 104
5.4 Results and Discussions 105
5.4.1 Results 105
5.4.2 Summary of main findings 106
5.4.3 General discussions 108
Efficacy of food visualisation 108
Effects of visual rendering quality 109
5.5 Limitations 109
5.6 Conclusion 110

6 Presentation of visual information near the mouth 111
6.1 HMD with a wide downward field of view 111
6.2 HMD that expands the downward FoV 112
6.2.1 System configuration of VR part 112

6.2.2 FoV of the prototype HMD of VR part 115
6.2.3 System configuration of VST part 116
6.2.4 Visual modulation function using machine learning 116
6.3 Experiment 119
6.3.1 Overview 119
6.3.2 Procedure 119
6.3.3 Questionnaire 121
6.3.4 Hypotheses 122
6.3.5 Participants 123
6.4 Results 123
6.4.1 Discussion 129
6.5 Conclusion 132

7 The effect of the avatar in the virtual environment changes on flavor perception 134
7.1 Presentation of the avatar without virtual mirrors 134
7.2 Experiment 1: Line Tracing Task 136
7.2.1 Overview 136
7.2.2 Participants 139
7.2.3 Procedure 139
7.2.4 Questionnaire 140
7.2.5 Hypotheses 141
7.2.6 Results 143
7.2.7 Discussion 144
7.3 Experiment 2: Escaping Task 148
7.3.1 Overview 148
7.3.2 Procedure 150
7.3.3 Hypotheses 151
7.3.4 Results 151
7.3.5 Discussion 154
7.4 Overall Discussion 156
7.5 Conclusion 158

8 Conclusion 160

8.1 Findings 160

8.2 Limitations and Future Work 162

8.2.1 Taste the foods we want to eat 162

8.2.2 Multimodal stimulus 163

8.2.3 Persistence of cross-modal effects 164

8.2.4 Visual modulation to the surrounding environment 164

8.2.5 Visual modulation at the moment the food is placed in the mouth 165

8.2.6 Effects of avatar display and changes in a physical appearance on multisensory flavor perception 166

8.3 Summary 168

Publication List 194

List of Figures

1.1 We propose three types of visual alteration methods to alter multisensory flavor perception. Food appearance) A method to change the appearance of food to different food. Environment) A method to change the appearance of the surrounding environment to the VE. Avatar appearance) A method to change the appearance of an avatar, a body used in the VE. 3

2.1 A study of multisensory flavor perception manipulation by texture [41]. Left) Smooth and rough textured sucrose. Right) Smooth and rough textured dishes. 11

2.2 Chewing Jockey manipulates multisensory flavor perception through auditory feedback [5]. 12

2.3 Multisensory flavor perception is altered by the combination of beverage flavor and color [45, 47] 13

2.4 A gustatory manipulation interface that changes multisensory flavor perception through LED coloring [9]. Top) overview. Bottom left) Method to separate the colored liquid from the beverage so as not to change the beverage's composition. Bottom right) in operation. 14

2.5 Gustatory manipulation interface that changes taste by multisensory presentation with LED coloring, olfactory display, and electric taste [10] 14

2.6 Experiments in which the shape and color of the tableware were varied [11]. Top) The shape of the dishes is changed. Bottom) The color of the dishes is changed. 15

2.7 A system that changes the size of a cookie held in hand [48]. Left) System operation. Top right) Calculation of the center of gravity of a cookie using an image acquired from a webcam. Bottom right) Resized hand and cookie. 16
2.8 Meta Cookie+ is a gustatory manipulation interface combining visual modulation by HMD and odor presentation by an olfactory display [4]. Top left) appearance of Meta Cookie+. Bottom left) experimental results on food taste. Top right) Meta Cookie+ in operation. Bottom right) schematic diagram of Meta Cookie+. . . 17
2.9 Food region detection using masked images can present texture according to deformation, even when cookies are missing or scattered [4]. 18
2.10 A gustatory manipulation interface that changes the appearance of tuna sushi into different sushi [14]. Top left) Tuna sushi changed to salmon sushi. Top right) Detection of tuna sushi region using convolutional neural network. Bottom left) Tuna being transformed into flatfish and different oily tuna. Bottom right) Recognition as a single food region even when the meal region is divided by a finger. 19
2.11 CycleGAN collects images with common features (called domains) and trains each generator and discriminator to transform each feature into each other [67]. 22
2.12 StarGAN for simultaneous conversion of multiple domains [73]. . . 23
2.13 Food images with appearance changes generated by CycleGAN [67] using a dataset of food images [74, 75]. 24
2.14 A dining experience with CAVE that changed the ambient appearance of the environment [76]. 26
2.15 Rounded surroundings environment make sweet foods taste even sweeter [17]. 27
2.16 The transparency switches between the VE and VST by switching the transparency depending on the angle of the head. Left) The virtual park environment. Right) The visible image from the mounted camera [18]. 27

2.17 Eating together experience in the VE. Left) realistic eating environment. Right) food superimposed on the VE and the user's hand [20]. 28

2.18 The user eats food with a red desk superimposed on the VE using red chrominance keying. Left) First-person view of the experience. Middle) External view of the experience. Right) Graphical model of the elements of the distributed reality [19]. 28

2.19 DeepLab results before and after input images and Conditional Random Fields (CRFs) when using the PASCAL VOC 2012 dataset [81]. 29

2.20 UECFood-100 images overlaid with segmentation masks annotated in UECFoodPix Complete [83]. 30

2.21 The result of SiamMask's tracking. Green frame) specifies a bounding box. Blue frame) around the target to be tracked in the first frame and automatically performs segmentation and tracking in the subsequent frames [85]. The frame box shows the tracking results of the Efficient Convolution Operators method [86]. 30

2.22 Panasonic HMD uses two fused lenses. 31

2.23 HMD with microlensarray for wide horizontal FoV [89]. 32

2.24 The forest scene is presented using an HMD that presents images in the peripheral vision. Left) Virtual scene rendered for Oculus Rift. Right) Scene displayed with the function to present images in the peripheral vision enabled [90]. 33

2.25 HMD with LC shutter installed at the bottom with changeable transparency, a keyboard of the real world can be displayed [94]. . 34

2.26 HMD with increased FoV by adding smartphones to the left, right, and bottom [95]. 34

2.27 Participants who were embodied in a virtual body that signifies super-intelligence. Einstein) improved their test scores. A) Einstein's virtual body. B) Normal virtual body. C) Participants wore an HMD, and their body movements were tracked [31]. 36

2.28 Participants with casual dark-skin avatars beat drums more rhythmically than in the other conditions. Left) Experimental conditions reflected in the mirror. A) White hand avatar. B) Casual dark-skinned avatar. C) Formal light-skinned avatar. Right) A participant wearing an HMD and a body-tracking suit, sitting on a stool and playing the drums [30]. 37

2.29 Participants strongly perceive virtual hand illusion [96] when using a realistic human hand. From left to right: realistic hand, toony hand, very toony hand, zombie hand, robot hand, wooden block [97]. 37

2.30 Full-body avatar is reflected in a virtual mirror. Left) Using CAVE. Right) Using HMD [103]. 38

2.31 Changing hand and forearm connectivity revealed that only the natural fully connected virtual limb condition elicited high ownership. A) Participants observed in first-person perspective through an HMD the virtual body in the same location and posture as the physical one. B) The visual appearances of the avatar' s right limb, Full-Limb: A standard full arm, Wire: A limb with a thin black rigid wire connecting the forearm and the hand, m-Wrist: A limb with a missing wrist, Plexiglass: A limb with a missing wrist with a Plexiglass panel placed in the blank space between the hand and the forearm [106]. 40

2.32 Participants' ability to judge distance is improved in the Inferior and Superior conditions with an increased vertical viewing angle. Typical) Normal vertical viewing angle. Inferior) Increased vertical viewing angle in the downward direction. Superior) Increased vertical viewing angle in the upward direction. Frame) White frame added outside the viewing angle, which is effective in improving the ability to judge distance [108]. 41

3.1 Left) Somen noodles in a cup with an AR marker (*N*ormal condition in the experiment). Right) Overlaid 3D model of ramen noodles (*O*verlay condition in the experiment). Although realistic, the model appearance does not change in accordance with the deformation of the actual food. It also causes an occlusion problem. 44

3.2 Results of the preliminary experiment. 46
3.3 Our GAN-based real time food-to-food translation system in action. Left) Input food images. Middle) User with a VST HMD experiencing vision-induced gustatory manipulation. Right) Examples of translated images. 48
3.4 System configuration of the GAN-based food-to-food translation system. The client module acquires an RGB image from the camera (a), sends it to the server (b), overlays the processed image on the video background (g), and presents the scene to the user (h). The server module receives the sent image (c), crops the center (d), translates it to another food image (e), and sends it back (f). 49
3.5 Left) Video background. Middle) schematic view of the image overlay. Right) corresponding translated image with radial gradient transparency. 50
3.6 Overall architecture of StarGAN [73] model, consisting of two modules: a discriminator D and a generator G. The discriminator D learns to distinguish whether the input image is fake or real and classifies it into the class it belongs to. The generator G learns to generate images that deceive D. 51
3.7 Results for various noodle conditions. The box plots in the upper and lower rows correspond to the results for **Q1** to **Q3** and **Q4** to **Q6**, respectively. 58
3.8 Left) The 3D model is static and it occludes the original food (somen noodles). Right) StarGAN dynamically generates a food image (fried noodles) that is visually coherent with the original food without occlusion. 60
3.9 Results of the accuracy of participants’ recognition of food. Before eating the food, participants answered on a VAS (0 and 100 being ‘strongly disagree’ and ‘strongly agree’, respectively). . . . 61
3.10 Results for rice conditions. The box plots in the upper and lower rows correspond to the results for **Q1** to **Q3** and **Q4** to **Q6**. . . . 62

4.1 Our GAN-based real-time food-to-food translation system in action. Left) User with a VST HMD experiencing vision-induced gustatory manipulation of Rice Conditions. Right bottom) Examples of Noodle Conditions. 73

4.2 A participant in a noodle condition. They ate food placed on a black table surrounded by a white wall while wearing HMDs. . . . 74

4.3 VAS scores in the noodle conditions. The box plots in the upper and lower rows correspond to the results for **Q1** and **Q2**, and **Q3** and **Q4**, respectively. The red crosses indicate the mean values, and the dots indicate the outliers. For example, the upper left graph, where "It tasted like . . . " and ". . . Somen noodles" intersect, shows the results of Q1, where "It tasted like somen noodles" was asked, and **Sn1** shows the VAS score of the first bite under **Sn** conditions. In addition, the entire **Sn1**–**Sn5** is denoted as **SnTaS** for identification purposes. 77

4.4 Relative change in VAS scores from the first to the fifth bite for each participant. Participants were divided into three groups based on the differences in VAS scores between the first and fifth bites: **Up** group (those increased by 10 points or more), **Down** group (those decreased by 10 points or more), and **Stay** group (those changed by less than 10 points). For example, the upper left graph, where "It tasted like. . . " and ". . . Somen noodles" intersect, shows the results of **SnTaS**, where "It tasted like somen noodles" was asked, and the *n*-th VAS score shows the VAS score of the *n*-th bite under Sn conditions. 78

4.5 VAS scores with regard to nationality and gender for **Q1** and **Q2** in the noodle conditions. Interactions are indicated to the right of result group names **SnTaS**, **FnTaS**, and **FnTaF**. The red crosses indicate the mean values, and the dots indicate the outliers. The upper graphs show the nationality classification, where Int indicates international participants and Jpn indicates Japanese participants. The lower graphs show the classification of gender, where F indicates female participants and M indicates male participants. For example, the upper left graph shows the values of **Sn1**–**Sn5** of **SnTaS** focusing on nationality. 80

4.6 VAS scores with regard to nationality and gender for **Q3** and **Q4** in the noodle conditions. Interactions are indicated to the right of result group names **SnTyS** and **FnTyS**. The red crosses indicate the mean values, and the dots indicate the outliers. The upper graphs show the nationality classification, where Int indicates international participants and Jpn indicates Japanese participants. The lower graphs show the classification of gender, where F indicates female participants and M indicates male participants. For example, the upper left graph shows the values of **Sn1**–**Sn5** of **SnTyS** focusing on nationality. 82

4.7 VAS scores in the rice conditions. The box plots in the upper and lower rows correspond to the results for **Q1** and **Q2**, and **Q3** and **Q4**, respectively. The red crosses indicate the mean values, and the dots indicate the outliers. For example, the upper left graph, where “It tasted like ...” and “...Steamed rice” intersect, shows the results of Q1, where “It tasted like steamed rice” was asked, and **Sr1** shows the VAS score of the first bite under **Sr** conditions. In addition, the entire **Sr1**–**Sr5** is denoted as **SrTaS** for identification purposes. 85

4.8 Relative change in VAS scores from the first to the fifth bite for each participant. Participants were divided into three groups based on the difference in VAS scores between the first and fifth bites: **Up** group (those increased by 10 points or more), **Down** group (those decreased by 10 points or more), and **Stay** group (those changed by less than 10 points). For example, the upper left graph, where "It tasted like ... " and "...Steamed rice" intersect, shows the results of **SrTaS**, where "It tasted like steamed rice" was asked, and the n-th VAS score shows the VAS score of the n-th bite under **Sr** conditions. 86

4.9 VAS scores with regard to nationality and gender for **Q1** and **Q2** in the rice conditions. Interactions are indicated to the right of result group names **CrTaS** and **CrTaC**. The red crosses indicate the mean values, and the dots indicate the outliers. The upper graphs show the nationality classification, where Int indicates international participants and Jpn indicates Japanese participants. The lower graphs show the classification of gender, where F indicates female participants and M indicates male participants. For example, the upper left graph shows the values of **Sn1**–**Sn5** of **SrTaS** focusing on nationality. 89

4.10 VAS scores with regard to nationality and gender for **Q3** and **Q4** in the rice conditions. Interactions are indicated to the right of result group names **CrTyS** and **CrTyC**. The red crosses indicate the mean values, and the dots indicate the outliers. The upper graphs show the nationality classification, where Int indicates international participants and Jpn indicates Japanese participants. The lower graphs show the classification of gender, where F indicates female participants and M indicates male participants. For example, the upper left graph shows the values of **Sn1**–**Sn5** of **SrTyS** focusing on nationality. 90

5.1 Flow of Ukemochi system, which consists of two modules – an Ukemochi client and server module. 101

5.2 Food scenes using Ukemochi in VEs. Top) Conditions for eating bread by hand. Bottom) Conditions for eating fried rice on a plate. **Hand**, **Plate**) The real-world environment for each meal condition. **Hn**, **Pn**) **None** condition without the real image. **Hv**, **Pv**) **VST** condition with the superimposed VIVE front camera image. **Hs**, **Ps**) **Seg** condition with only the food region cropped. 101

5.3 IPQ results. Top) Results for **Hand** condition. Bottom) Results of **Plate** condition. The red crosses indicate the mean values, and the dots indicate the outliers. [1 (strongly disagree) to 7 (strongly agree)]. 105

5.4 VAS results and the time to finish eating result. Top) Results for **Hand** condition. Bottom) Results of **Plate** condition. The red crosses indicate the mean values, and the dots indicate the outliers. The three graphs from the left show VAS results [0 (poor) to 100 (excellent)]. The rightmost graph shows the time to finish eating (s). 108

6.1 Schematic of the proposed HMD. A) Top view, B) Side view: yellow, display units; light blue, eyepiece; light green, fish eye camera; black lines, exterior of the HTC VIVE; orange lines, newly added parts. 113

6.2 Prototype HMD. A) Front view, B) Back view, C) Bottom display placement relative to the face. 113

6.3 Images displayed on the HMD and the mesh shape before and after mesh deformation. A1) Display image before mesh deformation captured by the fisheye camera. A2) Mesh shape before deformation. B1) Display image after mesh deformation captured by the fisheye camera. B2) Mesh shape after deformation. C) Correct pattern placed in front of the fisheye camera, each line represents a viewing angle in 5 [deg] increments. 114

6.4 Our HMD with a wide downward FoV. A) Side view. Blue: vertical FoV of the HTC VIVE. Part of the lower FoV is missing owing to the Liquid Crystal Displays (LCDs) housing. Orange: increased vertical FoV. Gray: vertical FoV of a typical human. B) The added LCD and Fresnel lens. C) Original FoV of the HTC VIVE (approx. 70×70 [deg]). The user's hand is visible but not the lower limbs and body. D) FoV of the prototype (approx. 70×130 [deg]). The user's lower limbs and body and the ground are all visible at the same time. 115

6.5 Our VST-HMD with a wide downward FoV. A) Side view. Blue: vertical FoV of the HTC VIVE. Part of the lower FoV is missing owing to the LCDs housing. Orange: increased vertical FoV. Green: vertical FoV of a VST–AR area. B) FoV of the prototype (approx. 70×100 [deg]). The upper and lower parts of the actual food are visible at the same time. 117

6.6 Viewing angle and video output for each condition. OffSn, OffFn) horizontal: approx. 70 [deg], vertical: approx. 50 [deg]. OnSn, OnFn) horizontal: approx. 70 [deg], vertical: approx. 100 [deg] (maximum viewing angle displayed)) 120

6.7 Diagram of the experiment. A: Side view, B: Red markings on the desk and wall used in the experiment, bowls, and chopsticks. . . . 121

6.8 The perceived taste of food results (Q1, Q2) [0 (strongly disagree) to 100 (strongly agree)]. 124

6.9 The recognized type of food results (Q3, Q4) [0 (strongly disagree) to 100 (strongly agree)]. 124

6.10 The perceived appearance of food results (Q5, Q6) [0 (strongly disagree) to 100 (strongly agree)]. 125

6.11 The perceived smell of food results (Q7, Q8) [0 (strongly disagree) to 100 (strongly agree)]. 125

6.12 The perceived food texture results (Q9, Q10) [0 (strongly disagree) to 100 (strongly agree)]. 126

6.13 Left) The ease of eating results (Q11) [0 (strongly disagree) to 100 (strongly agree)]. Right) The time results while eating [s] 126

6.14 The head pitch angle while eating results [deg]. Left) Maximum value. Middle) Maximum value. Right) Variance value. The head pitch angle is 0 [deg] when facing front and +90 [deg] when fully facing down. 127

6.15 The ease of eating results [0 (strongly disagree) to 100 (strongly agree)]. (Left) This experiment. Right) Preliminary experiment [142] 132

7.1 Our HMD with a wide downward FoV. A) Side view. Blue: vertical FoV of the HTC VIVE. Part of the lower FoV is missing owing to the LCDs housing. Orange: increased vertical FoV. Gray: vertical FoV of a typical human. B) The added LCD and Fresnel lens. C) Original FoV of the HTC VIVE (approx. 70×70 [deg]). The user's hand is visible but not the lower limbs and body. D) FoV of the prototype (approx. 70×130 [deg]). The user's lower limbs and body and the ground are all visible at the same time. 137

7.2 Course used in Experiment 1. A) Top view of the entire course (orange arrow indicates the direction of the course). B) Example user's view (participants were instructed to follow the purple line). C) Virtual cats to check the visibility of the downward FoV. . . . 138

7.3 Avatars used in Experiment 1. A) Male avatar. B) Female avatar. C) Spheres displayed on the hands. 139

7.4 IPQ results for Experiment 1 [1 (strongly disagree) to 7 (strongly agree)]. 143

7.5 IVBO results for Experiment 1 [1 (strongly disagree) to 7 (strongly agree)]. 143

7.6 Percentage of time spent on the purple line during the task (Score) for Experiment 1. 144

7.7 Course used in Experiment 2. Orange lines represent the course from the start to the goal. 2F) The second floor of the course. 1F) The first floor of the course. Door) A door that opens when touched by the avatar's hand for about 3 [s]. Fire) Avoidable tall fire on the course at about the head height. A) Avoidable short fire at about the shin height. B) Descending stairs from 2F to 1F. C) Obstacle near the head. D) Unavoidable tall fire at about the head height. 148

7.8 IPQ results and average head pitch angle at each location in Experiment 2. The five graphs from the left show IPQ results [1 (strongly disagree) to 7 (strongly agree)]. The two graphs from the right show the average head pitch angle at each location. Only important items corresponding to Fig. 7.7 A and B are shown (0 [deg] when facing front and +90 [deg] when fully looking down). . 152

7.9 Histogram of the frequency of occurrence of the head pitch angles at the location corresponding to Fig. 7.7 B in Experiment 2, ranked into 16 levels in 5 [deg] intervals (Orange: *OnH*; Blue: *OffH*). The vertical axis indicates the rank (the higher the more frequent) of each interval. The horizontal axis indicates the head pitch angle intervals (($X \leqq X_{raw} < X + 5$) at X [deg]). The head pitch angle is 0 [deg] when facing front and +90 [deg] when fully facing down. 152

1 Introduction

1.1 A method of presenting the multisensory flavors of food

Playwright George Bernard Shaw had the following aphorism in his play Man and Superman:
“There is no love sincerer than the love of food.”
Appetite is said to be one of the greatest desires of human beings, and is indispensable for maintaining life. The pleasure of being able to eat what you want, such as your favorite dishes and extravagant ingredients, is invaluable.

The question of how to cook safe and tasty food has long been a topic of human research. On the other hand, few engineering methods have been developed to determine how to manipulate people’s gustatory to make them perceive that they are eating a particular taste or food type. One reason is that it is difficult to define a base for gustatory, like the three primary colors for vision [1]. Therefore, it is challenging to present the gustatory of a particular food by pure taste stimulus combinations alone. The flavors perceived as tastes are not driven solely by pure gustatory stimuli. It is understood that taste perception is affected by the integration and interaction of different senses; in other words, all vision, hearing, olfaction, gustation, and tactile perception interact to perceive taste [2, 3]. Therefore, research is being conducted to present gustatory by applying the mechanism of multisensory integration in which stimuli from other senses are altered instead of altering pure taste stimuli [4, 5, 6].

In this doctoral dissertation, we define multisensory flavor perception [2] as the taste we perceive by integrating our various senses. We also define gustatory manipulation as altering multisensory flavor perception by presenting or changing sensory stimuli such as vision. In addision, the sensation we feel when we put food in our mouth (e.g., crunchiness) is called texture, but we define "food texture" to avoid confusion with the texture used in the image processing field.

In this doctoral thesis, we focus on visual information and manipulate gustation by modulating the vision. The first reason is that presenting and changing visual information methods are more established than other sensory presentations and can be modulated relatively easily using a head-mounted display (HMD). The second reason is that many studies have reported that color changes alter multisensory flavor perception [7, 8, 9, 10, 11, 12].

We hypothesized that users could manipulate gustation by visually changing the three elements visible to them while they eat (Figure 1.1):

1. The appearance of the food during the meal.

2. The appearance of the surrounding environments during the meal.

3. Visual information in the peri-personal space (Space around the body within reach of own hand) [13]. For example, the appearance of food near the mouth or the user's body.

To investigate these hypotheses, we have developed three different gustatory manipulation methods:

1. Gustatory manipulation interface that changes the appearance of food into the appearance of different food (Chap. 3 and Chap. 4).

2. Application that superimposes only the eating region on the Virtual Environment (VE) to investigate the effect of changing the appearance of the surroundings to the VE (Chap. 5).

3. HMD with an increased downward Field of View (FoV) to investigate the effect of visual information presented close to the mouth on the multisensory flavor perception (Chap. 6), and possibly to study the impact of changes in the avatar's body on the multisensory flavor perception (Chap. 7).

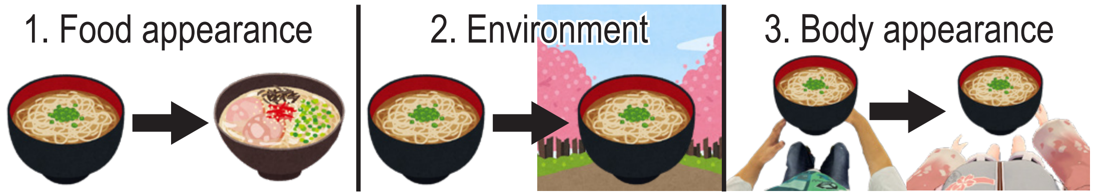


Figure 1.1: We propose three types of visual alteration methods to alter multisensory flavor perception. Food appearance) A method to change the appearance of food to different food. Environment) A method to change the appearance of the surrounding environment to the VE. Avatar appearance) A method to change the appearance of an avatar, a body used in the VE.

1.1.1 Change the appearance of food to different food intended

Augmented Reality (AR) which can visually change the appearance of food, can be used to change the type of food one is experiencing; this is done by superimposing three-dimensional (3D) models or images on the food. By combining an olfactory display with a variety of cookie images on top of plain cookies, Narumi *et al.* were able to transform the plain cookies into many different perceived flavors, for instance, chocolate [4]. Ueda and Okajima developed a gustatory manipulation interface (Magic Sushi) that uses machine learning to detect tuna sushi and change its appearance to salmon sushi, flatfish sushi, medium fatty tuna, and the fattiest portion tuna [14]. As a result, participants perceived more mouthfeel and oiliness in medium-fatty tuna and the fattiest portion of tuna sushi than tuna sushi. Current studies on how to manipulate gustation through vision involved mainly changing the image or color pattern of a single type of food, for example, sushis or cookies, significantly limiting its applicability and flexibility [4, 14]. There has been little research to determine whether these changes in multisensory flavor perception due to changes in food appearance persist throughout a meal and what attributes are most effective in humans.

Therefore, we investigated interactive systems that alter the appearance of one type of food into another and, more importantly, the impact of such a system on how food is experienced. We report how to use AR interfaces to manipulate the gustation and our manipulation interface using a generative adversarial network (GAN)-based real-time image-to-image translation (Chap. 3).

We also report the results of an investigation into the persistence of multisensory flavor perception changes by visual modulation using this gustatory manipulation interface, as well as the results of an investigation into whether the effect of multisensory flavor perception changes depending on the nationality and gender of the participants (Chap. 4).

1.2 Eating in the virtual world looks different from the real world

Several studies have reported that the multisensory flavor perception of participants changes in different eating environments. Meiselman *et al.* reported that participants rated restaurant meals higher than student refectories when eating the same foods [15]. However, the method in which participants actually move from place to place to change the appearance of the surroundings makes it difficult to control the experiment, and the places where meals are served are limited to realistic locations. We can more easily visually change the surrounding environment by using HMDs. In recent years, users have been able to experience VEs that differ significantly in appearance from Real Environment (RE) surroundings using inexpensive, general-purpose Virtual Reality (VR) HMDs. New VEs are increasing as companies and some users create and publish VEs with various world views, and some users spend most of their day in them.

On the other hand, it is difficult to perform the act of eating in the VE provided by the HMD because real food is not visible. For this reason, experiments have been conducted on eating in the VE by grabbing a bite-sized food item in a predetermined position [16] and drinking a beverage through a straw [17]. The problem is that eating without seeing the food decreases the ease of eating and reduces the presence (the sensation of being in the VE). Some studies use a Video See-Through (VST) HMD to acquire RE images and implement functions to change transparency [18] and chroma key compositing [19, 20] into the VE created by the researcher-self, thereby displaying food products while in the VE. However, these methods can only be used in specific VEs, and they must implement the functionality to use them in an existing VE.

Therefore, we have developed a food overlay application that allows existing VEs to influence flavor perception by changes in the appearance of the surrounding environment. We evaluate a proposed method of superimposing only the food segmentation images on the VEs using machine learning techniques to improve the ease of eating while maintaining a high presence (Chap. 5). We also report the results of an experiment measuring whether changes in the appearance of the surrounding environment affect the taste and flavor of food.

1.3 Alter the appearance of space close to our bodies

We investigated in Chap. 3, Chap. 4, and Chap. 5 the effects of changes in the appearance of the food and the appearance of the surrounding environment on multisensory flavor perception. In the course of these investigations, we found that existing HMDs have a limited vertical downward FoV, making it impossible to see food near the mouth. Two significant problems occur when food near the mouth is not visible. The first problem is the ease of eating when wearing the HMD. If the mouth area is not visible, it is difficult for the user to determine the relative position of the food to their mouth. The front camera of an HMD that acquires images of food is usually installed on the entire surface of the HMD. Therefore, there is a discrepancy between the position of the front camera and the actual position of the user's eyes. Because the human face is structured so that the mouth is directly below the eyes, the user may misidentify the mouth as being directly below the front camera position. These causes make it difficult for the user to eat with the HMD in place. Several experimental participants in Chap. 3, Chap. 4, and Chap. 5 also reported that the lack of visibility near the mouth inhibited their ease of eating.

The second problem is the inability to alter the food appearance near the mouth. In the experiments in Chap. 3 and Chap. 4, multisensory flavor perception was altered by a cross-modal effect of visual modulation that changed the food appearance. However, it is impossible to display the altered food appearance when it is placed in the mouth the moment the user perceives the taste. We hypothesize that limited downward FoV suppresses the cross-modal effects of visual modulation on gustation.

Therefore, we have developed a novel VST-HMD with an increased downward FoV by adding two new optics to the bottom of an existing HMD (Chap. 6). We report the results of an experiment in which we investigated whether displaying food in the downward FoV improves the ease of eating while wearing the HMD and the amount of cross-modal effect on gustation.

Multisensory integration occurs more frequently in peripersonal space, the space within human reach [21, 22]. In our daily lives, our bodies are always present in the peripersonal space. The virtual body, an avatar, is displayed in the VE and follows and synchronizes with the user's movements to improve presence. Sense of embodiment (SoE), including the sense of self-location (SoSL), the sense of agency (SoA), and the sense of body ownership (SoBO) represent the avatar's sense of being myself, and the higher these are, the better the presence [23]. However, displaying avatars with existing HMDs with limited downward FoV is difficult. When the user is facing forward, existing HMDs can only display the forearm from the avatar's elbow to the hand, and it isn't easy to display avatars other than the hand unless the head is turned down. In VR research using avatars, the avatar's body is reflected in a virtual mirror installed in the VE to make the user see the avatar's body [24, 25, 26]. Experiments using virtual mirrors have reported a positive effect on performance using full-bodied humanoid avatars [27], but no positive effect occurred in experiments without [28]. Furthermore, the multisensory integration model proposed by Ernst *et al.* shows that the higher the likelihood of sensory information, the more multisensory integration is promoted [29]. We expect that first-person avatar body displays will improve the likelihood of visual information.

Furthermore, several studies have reported that changing the appearance of an avatar can change the user's behavior and mental state and improve the user's abilities [30, 31]. The changes in the avatar's body affect the human psychological state and behavior, which is called the Proteus effect [32] and is said to depend on the user's impression or preconception of the avatar. We also hypothesize that changes in the avatar's body may influence multisensory flavor perception. For example, when using an avatar of a rich person, food may taste more expensive than everyday meals.

We believe that displaying the avatar's body in peripersonal space is necessary to study multisensory flavor perception using avatar. We surmise that setting up a mirror in front of you and eating while checking how you are eating your food is unnatural and interferes with multisensory integration. We thought that by displaying the avatar using the HMD with an increased downward FoV created in Chap. 6, we might see the avatar's body in the first-person perspective, except for the avatar's hands. As a preliminary step in examining the effect of changes in avatar body appearance on multisensory flavor perception, it is necessary to examine the effectiveness of the avatar's body display for downward FoV. Therefore, we develop an avatar display method using the HMD with an increased downward FoV and report the results of our investigation into the effectiveness of the avatar display for downward FoV (Chap. 7).

1.4 Contribution

In this doctoral dissertation, we developed and demonstrated the effectiveness of three gustatory manipulation methods to investigate the impact of visual changes on multisensory flavor perception by altering the appearance of food and environments through AR and VR.

Contributions are highlighted as follows:

1. We present a more flexible vision-induced gustatory manipulation system using GAN-based real-time image-to-image translation. Our system flexibly supports multiple types of target food, which changes its appearance dynamically and interactively in accordance with the deformation of the original food. Our experimental results confirmed that participants more strongly perceived the taste of the target food solely upon GAN-based visual modulation (Chap. 3).

2. We report that vision-induced gustatory manipulation is persistent in many participants for several times a mouthful of food. Their persistent gustatory changes are divided into three groups: those in which the intensity of the gustatory change gradually increased, those in which it gradually decreased,

and those in which it did not fluctuate, each with about the same number of participants (Chap. 4).

3. We also report that visually induced gustatory manipulation is affected differently depending on the participant's attributes (gender and nationality). Still, those who are less familiar with the original and target types of food may have a stronger effect (Chap. 4).

4. We investigate the effects of visually altering not only the appearance of the food but also the surrounding virtual environmental appearance. We propose a method that enables users to eat while maintaining the high presence of existing VEs. We show that superimposing only the food segmentation image on the VE gives a high presence and maintains ease of eating. However, our experiments examining the effects of changes in an environmental appearance on taste and flavor found no statistical differences (Chap. 5).

5. We argue that existing HMDs have a limited downward FoV and cannot present images near the mouth. We developed a VST-HMD with a wide downward FoV to solve this problem. We investigated the effects of video presentation near the mouth on taste change and palatability and found no significant differences (Chap. 6).

6. We report that gustatory manipulation using visual modulation can enhance not only the perception of taste and type of food but also the perception of smell and food texture of the visually presented food (Chap. 6).

7. We report a positive effect on presence and SoSL using the avatar display for the downward FoV, with more natural head movements. However, the effects on SoBO and SoA were limited due to the perceived misalignment between real and virtual bodies. These results predict using avatars will advance the study of multisensory flavor perception. These results also suggest the need to improve the accuracy of body tracking and explore avatar representations that are acceptable as one's own body (Chap. 7).

2 Related Works

2.1 Mechanism of human multisensory flavor perception

The first method we can use to describe the taste of food in an engineering way is to present a combination of basic tastes for the sense of taste (gustatory sensations). There are theories that the four basic tastes are sweet, salty, sour, and bitter, and others that the five basic tastes, including umami, are the five basic tastes. However, there are several definitions of basic tastes, such as the research report that taste is a continuum and that there is no such thing as a basic taste [1] and the research report that there is no single receptor site for sweet and bitter tastes [33]. In addition, due to the influence of painful stimuli such as pungency and the fact that the threshold for perceiving taste (e.g., sweetness) differs among individuals [34], a method for presenting taste by itself has yet to be established.

On the other hand, research has revealed that our gustatory sensations are affected by gustatory stimuli and by the other senses [2]. The taste we perceive is perceived as a multisensory flavor perception that integrates these sensory stimuli. In other words, multisensory flavor perception is strongly influenced by stimuli from the tongue's taste cells and smells, sounds, and the surrounding environment. For example, to recognize that we have eaten an apple, we integrate many senses (multisensory integration), such as the apple's red color and glossy images (vision), refreshing smell (olfaction), hard food texture (tactile), crunchy sound (hearing), and sweet taste (gustation).

Thus, humans integrate and process information obtained through various senses, including the five senses, in the brain to form perception and recognition. The general term for the cognitive process of integrating multisensory information is called integrated perception/cognition, and it has been suggested that information from different senses strongly influences each other [35].

Perceptual characteristics relevant to the gustatory presentation using integrated cognition include multisensory integration, in which stimuli input to different sensory modalities are perceived and recognized as a single event. Ernst *et al.* proposed a multisensory integration model using Bayes' theorem to explain the nature of multisensory integration [29, 36]. This multisensory integration model [29] uses the relationships among sensory information obtained in daily life as a prior distribution. In other words, multisensory integration in gustation is influenced by the current food experience to date.

Therefore, if a food is perceived to have a different appearance than the food currently being eaten, it may be perceived as tasting differently. This type of perception is called cross-modal perception, in which the perception of one sense is affected by stimuli to other senses that are received simultaneously.

As described above, it is difficult to express the presence of gustation with pure taste stimuli alone. However, research is being conducted to induce a cross-modal effect through the multisensory presentation by utilizing the integrated cognitive characteristics of humans and presenting information from various senses to present gustation.

2.2 Gustatory manipulation by multisensory presentation

In this section, we introduce gustatory manipulation techniques using various sensory presentations. Olfaction is the sense that has the greatest influence on multisensory flavor perception [37]. For example, the flavor of food is heavily affected by signals from the taste buds associated with smell [37, 38, 39]. For example, Murphy and Cain reported that their experimental participants tasted ethyl acetate, a flavor component of peaches and pears, despite that it is actually tasteless [40]. However, they also reported that up to 80% of the taste disappeared when their nose was blocked. Stevenson *et al.* found that the sweet taste of sucrose increased and the sour taste of citric acid decreased when presented with the odor of sweet caramel [39]. Various gustatory displays make use of these

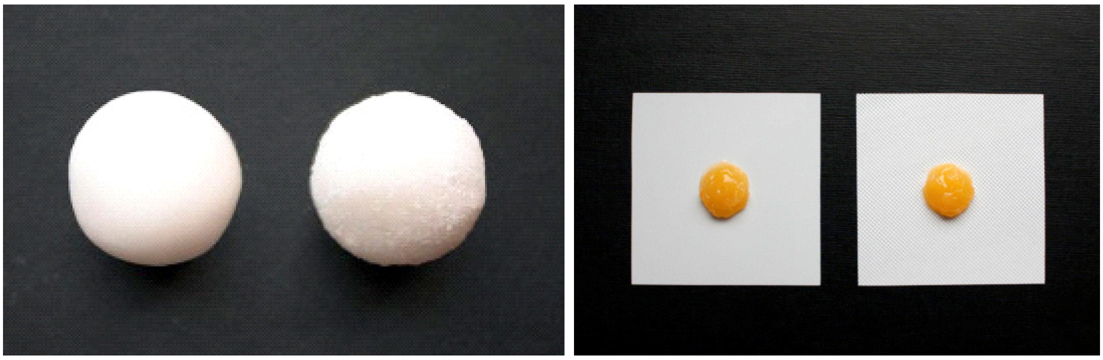


Figure 2.1: A study of multisensory flavor perception manipulation by texture [41]. Left) Smooth and rough textured sucrose. Right) Smooth and rough textured dishes.

multimodal and cross-modal effects of olfaction [4, 10].

When eaten, the food texture also produces tactile and auditory sensations, and thus gustatory displays that make use of these sensations are emerging [5, 6]. Besides, the tactile and the sense of hearing influence the multisensory flavor perception. As an example of the influence of the tactile on the guatation, Slocombe *et al.* reported that the acidity perceived by participants was stronger in rough foods than in smooth foods (Figure 2.1) [41]. Iwata *et al.* have developed a food simulator as a gustatory manipulation interface that can reproduce food texture by measuring and reproducing the temporal changes in force when chewing food [6]. As for the effect of hearing on taste, Zampini and Spence found that when a high-pass filtered chewing sound or white noise was played in the ear while chewing deep-fried potato chips, the chips felt crispier and fresher [42]. Koizumi *et al.* developed Chewing Jockey as a gustatory manipulation interface based on this phenomenon (Figure 2.2) [5]. Chewing Jockey successfully amplifies the crunchiness of potato chips, the thickness of cookies, and the stickiness of daifuku by processing the subject's chewing sound and playing it back through earphones. Spence *et al.* found that bacon or egg flavor was more strongly perceived when the sound of bacon cooking or chicken clucking was played while eating bacon egg-flavored ice cream [43]. In the field of electric taste, Miyashita developed gustatory displays that have been using ion electrophoresis to present tastes by individually suppressing the five basic tastes contained in five gels [44].

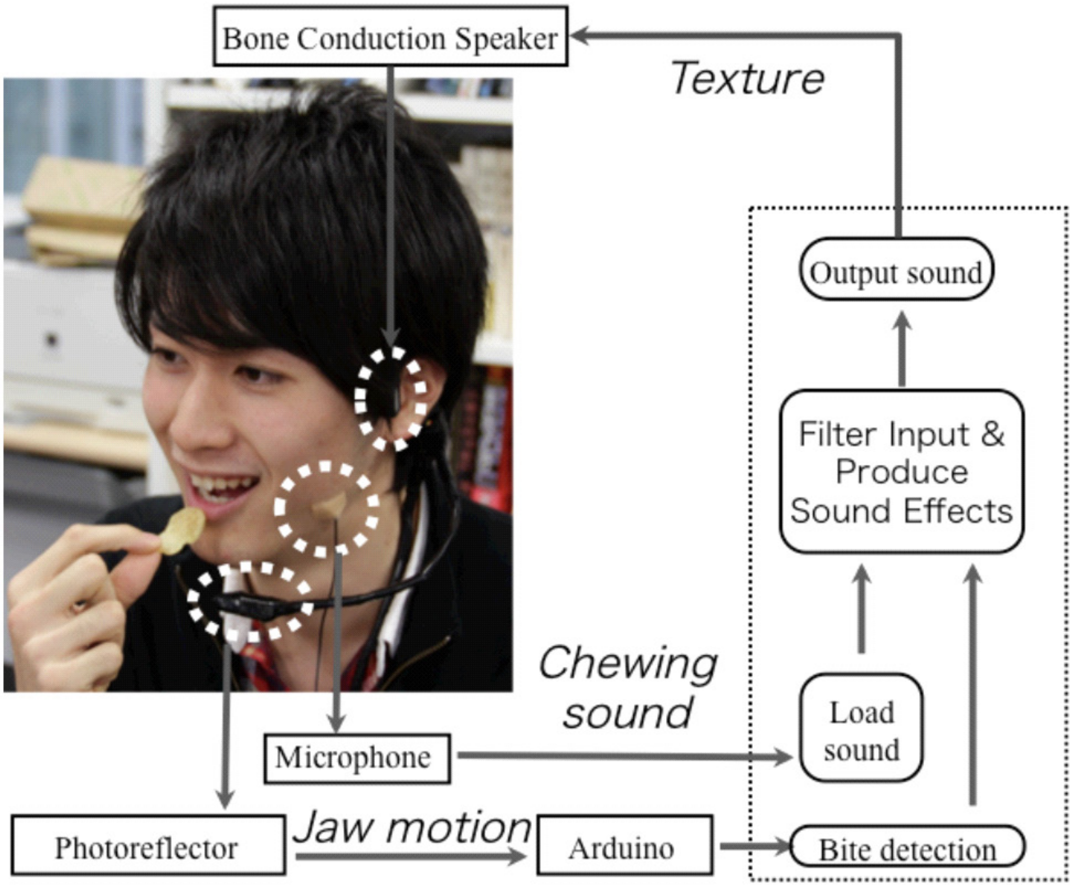


Figure 2.2: Chewing Jockey manipulates multisensory flavor perception through auditory feedback [5].

Vision is known to affect taste [45] as put by van der Laan *et al.* 'The first taste is always with the eyes' [46]. Additionally, visual modulation of food affects the perception of food types. Some studies have successfully changed how the flavor of food is perceived by altering its color [7, 8, 9, 10, 11, 12]. For example, by changing the color of wine Morrot *et al.* demonstrated that they were able to make sommeliers believe that white wine tasted like red wine [8]. Stillman measured how accurately the average participant with no gustatory training could discriminate taste when drinking raspberry- and orange-flavored colored beverages (colorless, red, orange, and green) [47]. The experiment results revealed that participants could accurately identify the raspberry flavor when they drank

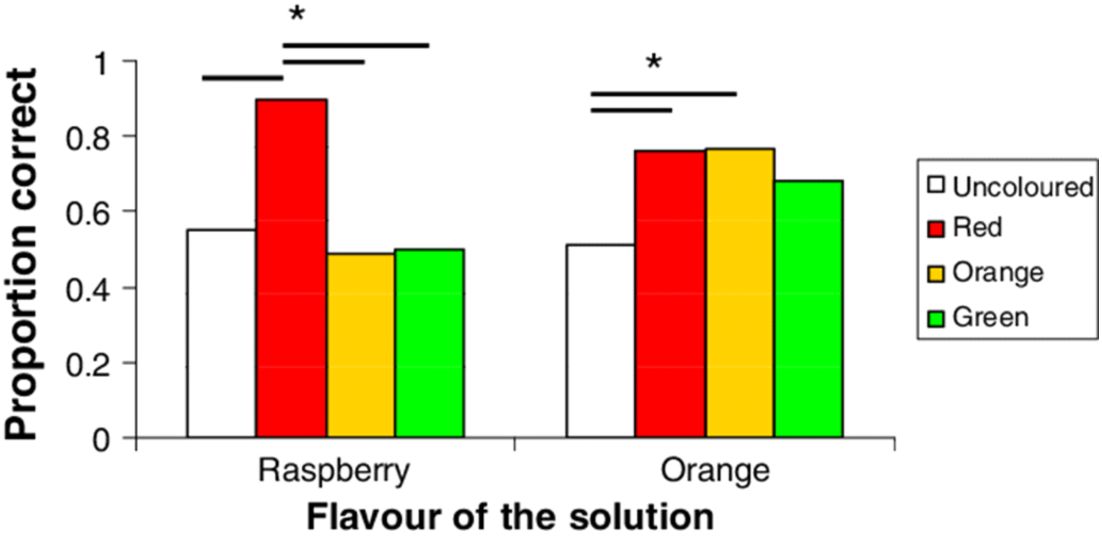


Figure 2.3: Multisensory flavor perception is altered by the combination of beverage flavor and color [45, 47]

a red-colored beverage (Figure 2.3). They also tended to identify the orange flavor more accurately in the colored drinks than in the colorless liquids. Narumi *et al.* used LEDs to color beverages and change the same flavored beverage to a different flavor (Figure 2.4) [9]. Additionally, Ranasinghe *et al.* changed the perceived flavor of water in a cocktail glass by combining color, smell, and electrical taste stimuli to the water (Figure 2.5) [10]. Zampini *et al.* conducted an experiment in which participants drank orange- and blackcurrant-flavored beverages that were colored in various ways and found that proper visual presentation with flavor-matching coloration helped them identify the taste accurately [7].

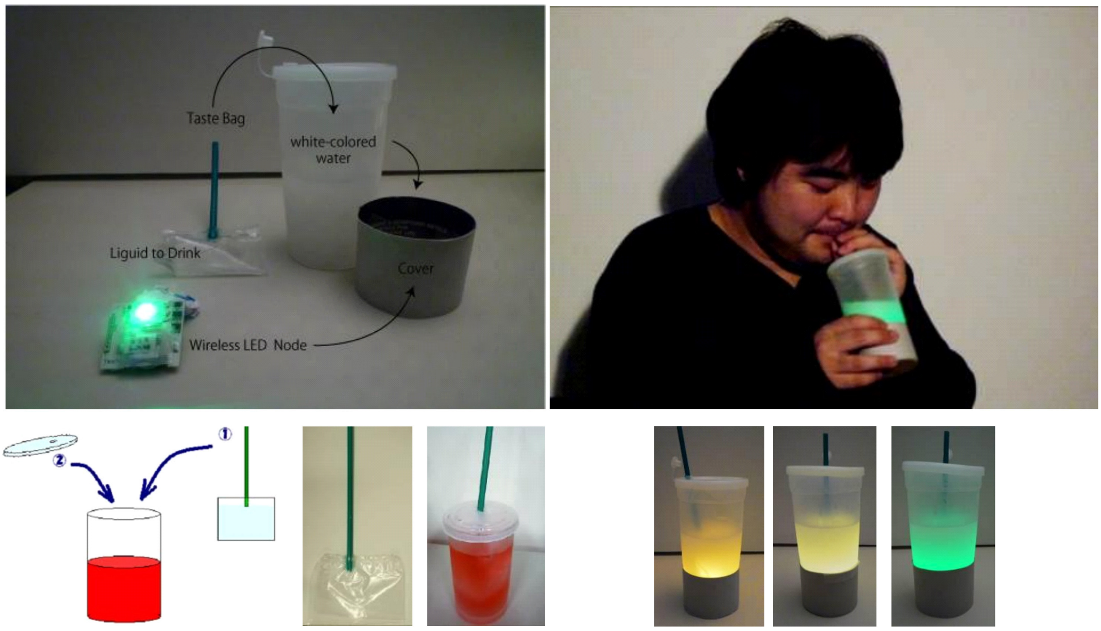


Figure 2.4: A gustatory manipulation interface that changes multisensory flavor perception through LED coloring [9]. Top) overview. Bottom left) Method to separate the colored liquid from the beverage so as not to change the beverage's composition. Bottom right) in operation.

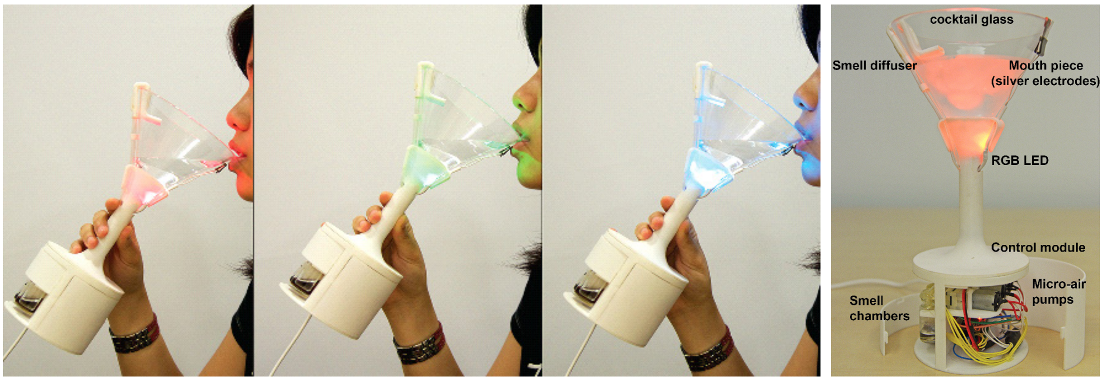


Figure 2.5: Gustatory manipulation interface that changes taste by multisensory presentation with LED coloring, olfactory display, and electric taste [10]

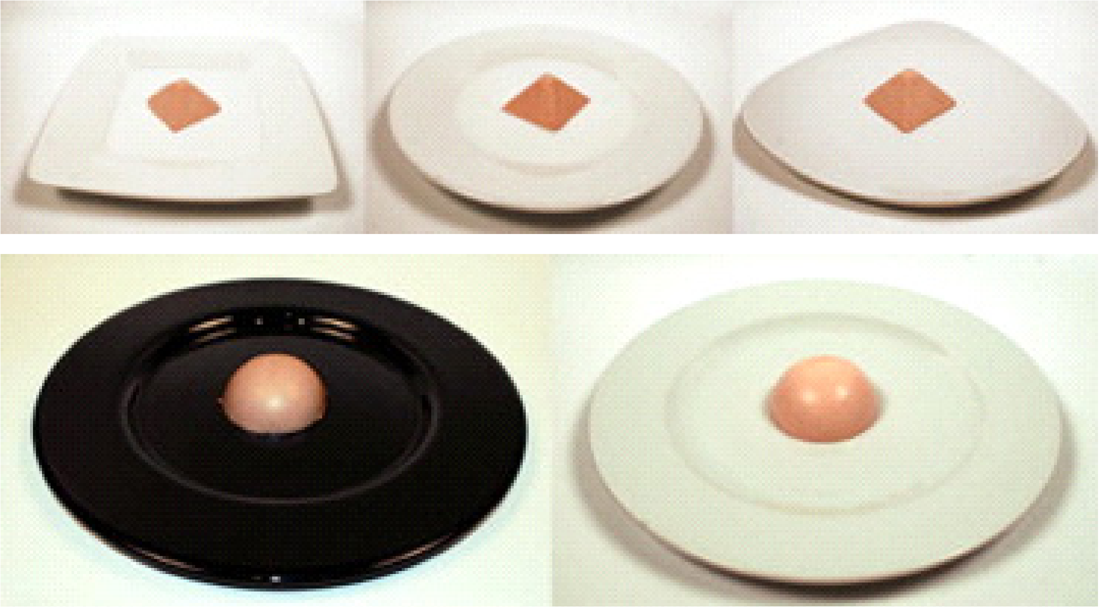


Figure 2.6: Experiments in which the shape and color of the tableware were varied [11]. Top) The shape of the dishes is changed. Bottom) The color of the dishes is changed.

Piqueras-Fiszman *et al.* found that participants perceived the sweetness of strawberry-flavored mousse more strongly and preferred it more when eating it on a black plate than when using a white plate (Figure 2.6) [11]. Shankar *et al.* [12] showed that the lower the degree of inconsistency between the participants' expected flavor and the color combination of the beverage, the more it affected their perception. Besides, they showed that food perception manipulation could be performed even if participants are explicitly told that color is an uninformative cue. They know that they are eating visually altered food.

Besides the perception of gustation, vision also influences the feeling of satiety. The feeling of satiety felt by the participant can be manipulated in several methods, e.g., changing the color of the cup [49], using a bowl that automatically and continuously refills with soup to prevent the food from visually diminishing [50], using the Delboeuf illusion to misrepresent the size of the food [51], using AR to change the size of food (Figure 2.7) [48].

AR can also be used to change the type of food one is experiencing; this is

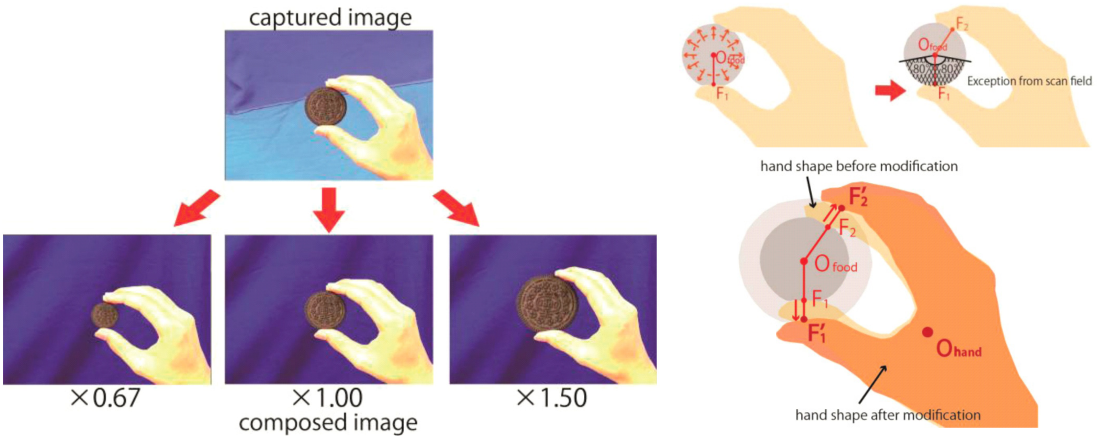

captured image
×0.67
×1.00
composed image
×1.50
O food
F₁
F₂
80
80
Exception from scan field
hand shape before modification
F′₂
F₂
O food
F₁
F′₁
O hand
hand shape after modification

Figure 2.7: A system that changes the size of a cookie held in hand [48]. Left) System operation. Top right) Calculation of the center of gravity of a cookie using an image acquired from a webcam. Bottom right) Resized hand and cookie.

done by superimposing 3D models or images on the food. By combining an olfactory display with a variety of cookie images on top of plain cookies, Narumi *et al.* were able to transform the plain cookies into many different perceived flavors, for instance, chocolate [4]. Ueda and Okajima developed a gustatory manipulation interface that uses machine learning to detect tuna sushi and change its appearance to salmon sushi, flatfish sushi, the medium fatty tuna, and the fattiest portion of tuna (Figure 2.10) [14]. As a result, participants perceived more mouthfeel and oiliness in medium-fatty tuna and the fattiest portion of tuna sushi than tuna sushi. The studies mentioned above show that it is feasible to change the type of food that people believe they are consuming by altering its appearance. However, current studies on vision-induced gustatory manipulation have only experimented with a single type of food, such as cookies or sushi, considerably limiting its applicability and flexibility. To our knowledge, no studies on vision-induced gustatory manipulation have examined the effects between different types of food. Not much is known as to whether such manipulation is possible and to what extent.

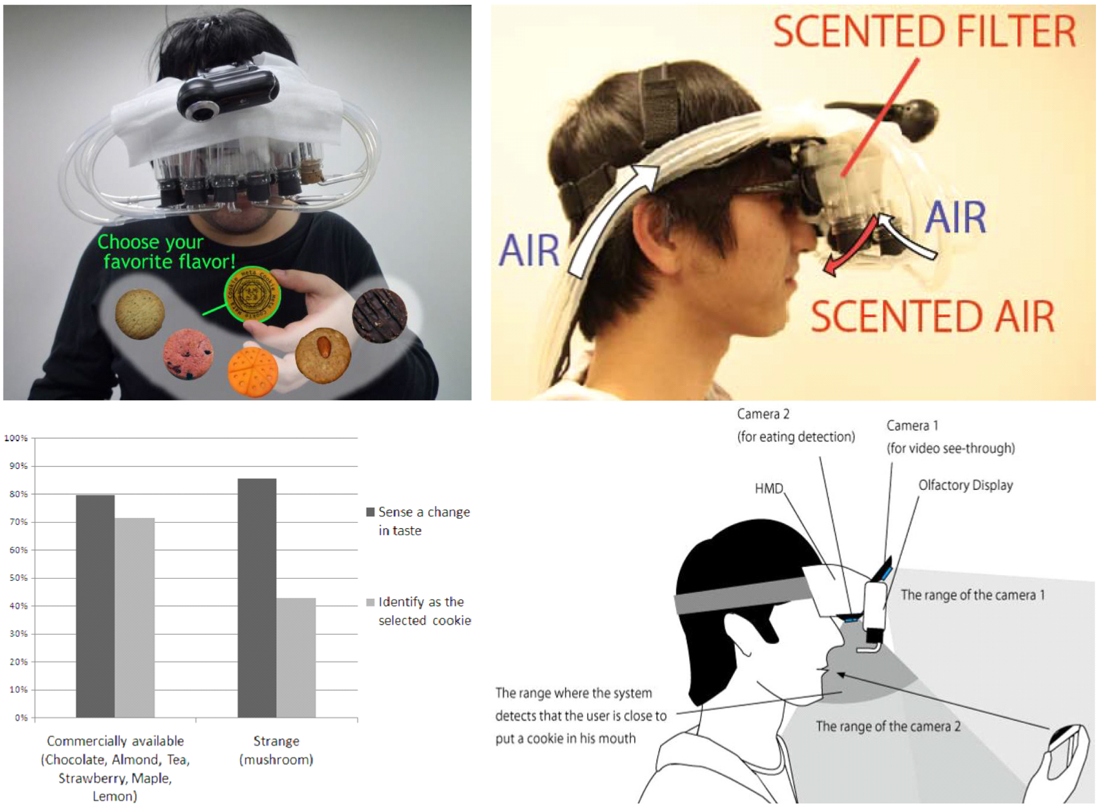

Choose your
favorite flavor!
SCENTED FILTER
AIR
AIR
SCENTED AIR
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Commercially available (Chocolate, Almond, Tea, Strawberry, Maple, Lemon)
Strange (mushroom)
■ Sense a change in taste
■ Identify as the selected cookie
Camera 2 (for eating detection)
Camera 1 (for video see-through)
HMD
Olfactory Display
The range of the camera 1
The range where the system detects that the user is close to put a cookie in his mouth
The range of the camera 2

Figure 2.8: Meta Cookie+ is a gustatory manipulation interface combining visual modulation by HMD and odor presentation by an olfactory display [4]. Top left) appearance of Meta Cookie+. Bottom left) experimental results on food taste. Top right) Meta Cookie+ in operation. Bottom right) schematic diagram of Meta Cookie+.

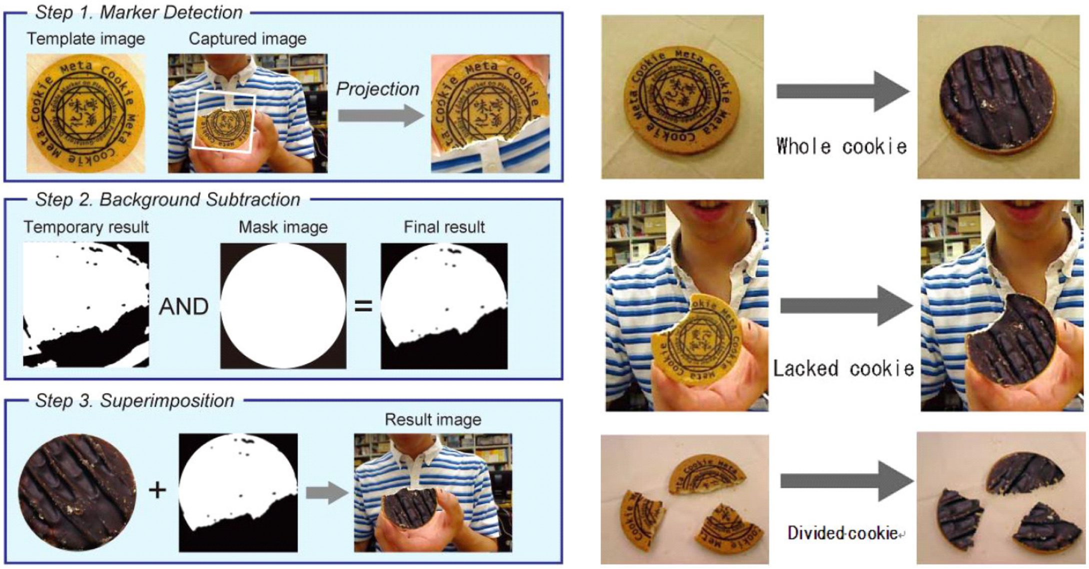


Figure 2.9: Food region detection using masked images can present texture according to deformation, even when cookies are missing or scattered [4].

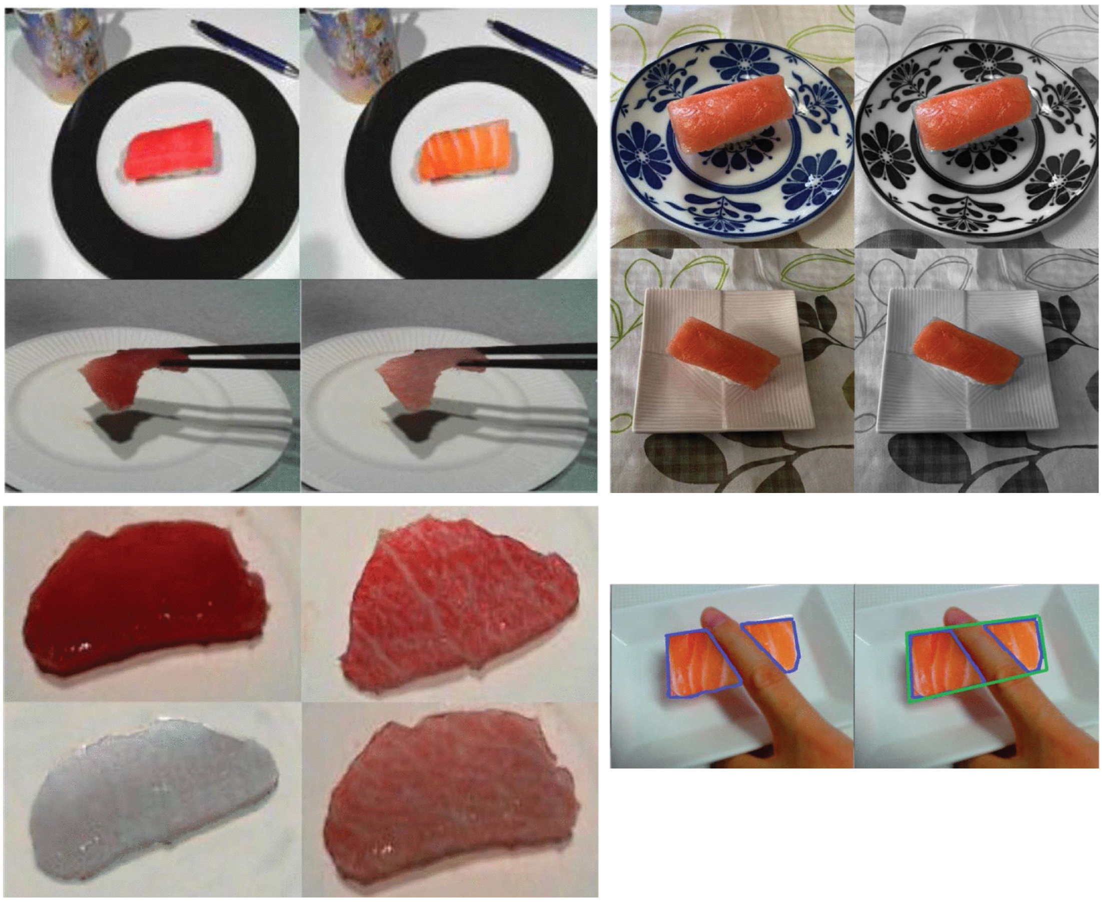


Figure 2.10: A gustatory manipulation interface that changes the appearance of tuna sushi into different sushi [14]. Top left) Tuna sushi changed to salmon sushi. Top right) Detection of tuna sushi region using convolutional neural network. Bottom left) Tuna being transformed into flatfish and different oily tuna. Bottom right) Recognition as a single food region even when the meal region is divided by a finger.

Besides, if it is possible to change the perception of the types of food actually eaten, it may improve the quality of life (QoL) of people with dietary restrictions. There are circumstances in which people cannot eat what they want for reasons such as dietary restrictions, food allergies, and religious restrictions. In such cases, their QoL will significantly decline [52, 53].

For example, a survey by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) revealed that 2–4% (about 300-600,000) of elementary, junior high, and high school students in Japan have some food allergy [54]. One of the problems faced by patients diagnosed with a food allergy is a decrease in dietary QoL due to the time and effort required to remove the causative food and the various processed foods it contains, as well as anxiety about the possibility of food allergy symptoms [55]. It is also known that the QoL of parents of children diagnosed with food allergies declines [56, 57]. Furthermore, in addition to food allergy patients, for example, Crohn's disease patients are known to have reduced QoL due to symptoms and dietary restrictions [58]. In addition, there are many situations where dietary restrictions are necessary, such as dieting for weight loss, disease-fighting, and religious dietary restrictions. When men and women aged 20 years and older were surveyed about their dieting experience, 66.9% reported that they were currently dieting or had dieted in the pastt [*]. These results suggest that the inability to eat what they want to eat reduces the QoL of a large number of people.

In Chap. 3, we report in detail a user study on an AR system that overlays a 3D model of a target food over a different type of original food and its successful experimental results. We then summarize the identified problems of 3D model-based systems and describe in detail the implementation of and the main user study on the more flexible GAN-based AR system, which changes the appearance of one type of food into another in real time.

In Chap. 4, we provide a detailed examination of the effects of vision-induced gustatory manipulation revealed in Chap. 3. For a gustatory display to be used in daily eating, the effect of multisensory flavor perception presentation must persist throughout the meal. Therefore, we report the results of measuring flavor per-

[*] Health awareness survey, https://www.kirinholdings.com/jp/newsroom/release/2008/0303_01.html, last accessed March 17, 2023.

ception with each bite to clarify the persistence and change in multisensory flavor perception due to visual modification. Besides, the formation of cross-modal associations to the gustation differs depending on cultural differences such as the nationality and gender of the participants [59, 60, 61]. We compare the results of experiments on groups of participants of different nationalities and genders and investigated the effectiveness of our vision-induced gustatory manipulation on participants with varying food cultures.

We need to generate different food images from actual food images to develop the vision-induced gustatory manipulation for these experiments. Therefore, we develop vision-induced gustatory manipulation using GANs, which have attracted attention for their ability to produce high-quality images.

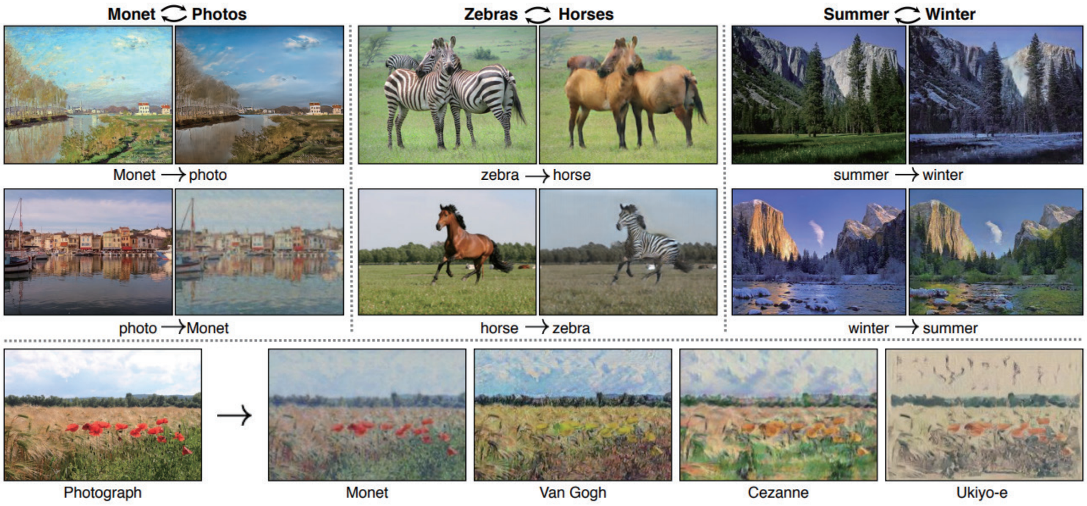


Figure 2.11: CycleGAN collects images with common features (called domains) and trains each generator and discriminator to transform each feature into each other [67].

2.3 Food image generation using generative adversarial networks

GANs [62, 63] have achieved impressive results in image generation [64, 65], image-to-image translation [66, 67, 68, 69], text-to-image translation [70, 71], and super-resolution [72] (Figure 2.11). GANs are composed of a generator and a discriminator. The discriminator learns to distinguish generated fake samples from real ones, while the generator learns to generate fake samples that are indistinguishable from the real ones. In this work, we also employ an adversarial loss to learn the mapping to make the translated images as realistic as possible.

Recent work has resulted in excellent image-to-image translation [66, 67, 68]. For example, pix2pix [66] uses paired images and learns this task by supervised learning. However, since it incorporates an L1 loss into an adversarial loss, pix2pix [66] has a problem that it requires paired data samples. To solve this problem, unpaired image-to-image translation frameworks [67, 68, 69] have been proposed. For example, CycleGAN [67] and DiscoGAN [68] make unsupervised translation possible by utilizing a cycle consistency loss between input and trans-

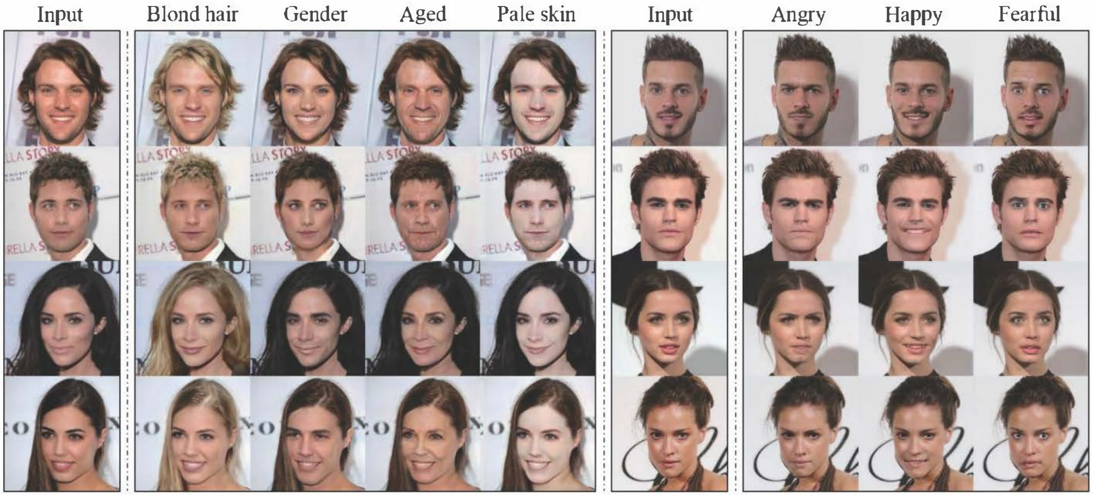


Figure 2.12: StarGAN for simultaneous conversion of multiple domains [73].

lated samples. However, all these methods are able to learn the transformation between only two domains at the same time. Therefore, their approaches require learning different transformation models separately to support more than two domains. In addition, this also limits the scalability to make many transformation models. To solve this problem, StarGAN [73], which enables the translation of multiple domains at the same time (Figure 2.12). StarGAN [73] imposes tasks on the discriminator not only to identify a fake or real image but also to identify which class a translated image belongs to using an auxiliary classifier. In Chap. 3 and Chap. 4, we employ StarGAN [73] so that multiple food transformations can be learned at the same time (Figure 2.13).

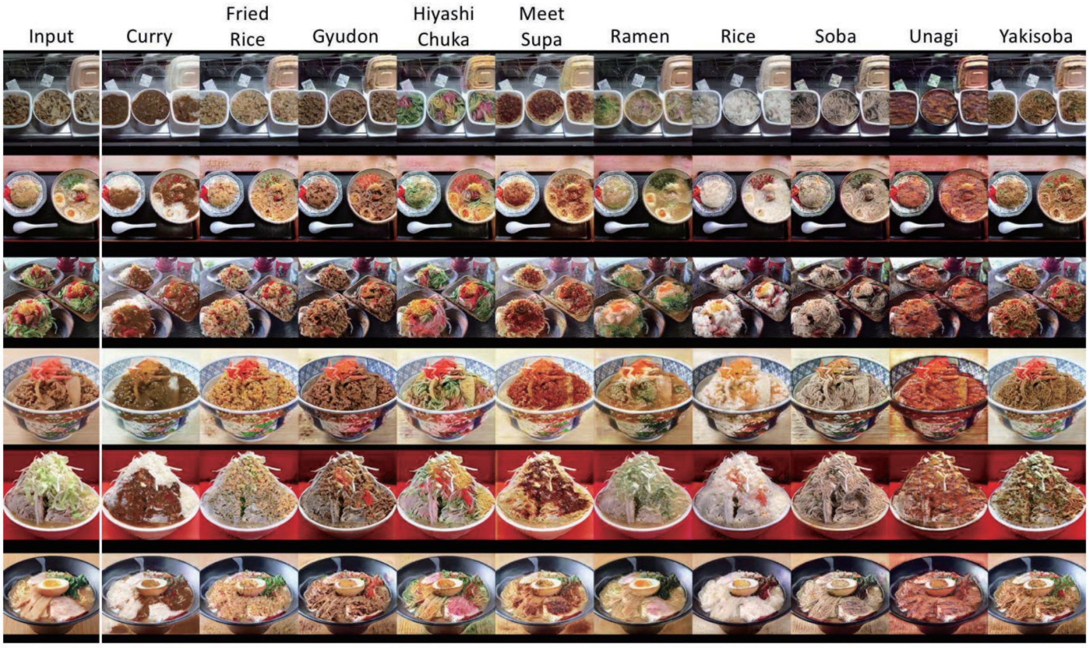


Figure 2.13: Food images with appearance changes generated by CycleGAN [67] using a dataset of food images [74, 75].

2.4 Eating while wearing HMD in the VE

We believe that it is possible to alter multisensory flavor perception not only by changing the appearance of food but also by changing the appearance of the environment surrounding food in the VE and altering visual information. For example, several studies have shown that changing the environment around the user can change the evaluation of meals. Meiselman *et al.* showed that differences in the surrounding environment could alter the evaluation of a meal [15]. Dionisio *et al.* used a Cave Automatic Virtual Environment (CAVE) for visual presentation to provide a highly immersive meal in the VE (Figure 2.14) [76]. An HMD can change the surrounding environment with high presence. However, the presentation of VE images alone does not allow the visibility of the real food.

Bouba/Kiki effect [77] investigated the correspondence between speech and the visual shape of objects. In multisensory flavor perception studies, spiky shapes (Kiki) were associated with sourness and rounded shapes (Bouba) with sweetness in both foods and beverages [78, 79]. Cornelio *et al.* found that rounding the actual sweet food shape in the VE increases the sweetness [16]. They also reported that VE under blue lighting decreased the sweetness of neutral-tasting foods. The experimenter placed the food in the same position as the 3D model so that the participant could grasp the food, although the food was not tracked. Chen *et al.* have made beverages taste sweeter in a rounded virtual space (Figure 2.15) [17]. Although participants could not see the beverage in their hands, they could drink it by relying on the sensation in their hands because the beverage had a straw attached. These studies suggest that changes in the surrounding environment can affect multisensory flavor perception. However, these experimental methods can only be applied to bite-sized foods and beverages, and it is challenging to use foods in containers because of the lack of visibility and tracking of the food.

In order to use a variety of foods in an experiment, it is necessary to display VST images on the VE. Korsgaard *et al.* proposed a method to switch between VE and VST images depending on the head orientation (Figure 2.16) [18]. For example, when the user looks down, the food on the desk is visible; the VE is visible when the participant looks forward. The VST image and the VR space are seamlessly switched between displays by adjusting the blending ratio. However, since all the real objects are visible, the sense of presence is significantly reduced.

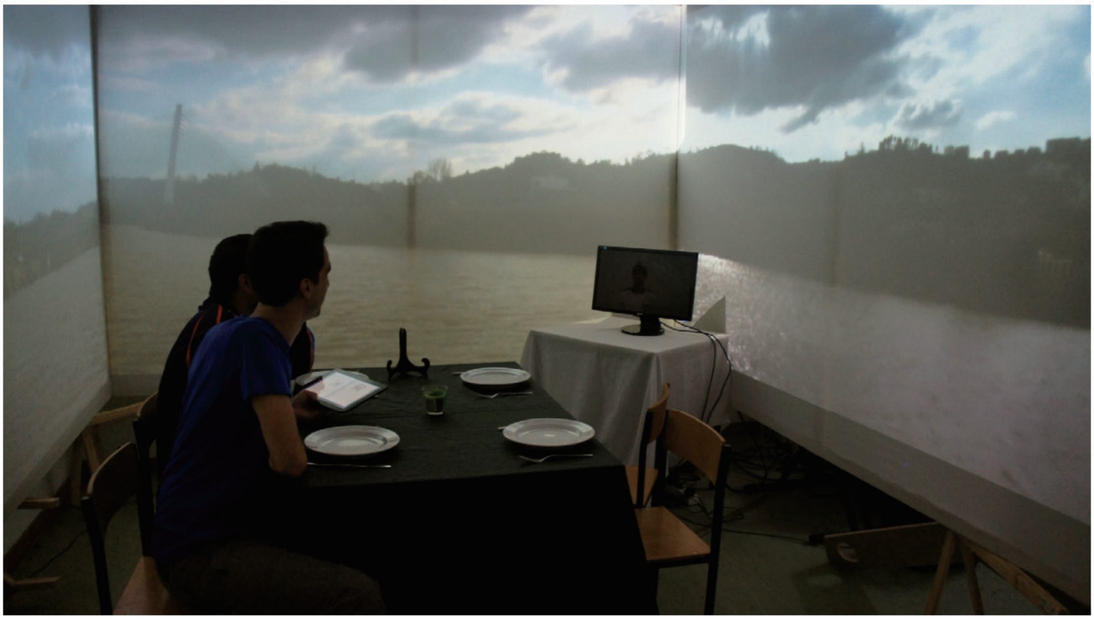


Figure 2.14: A dining experience with CAVE that changed the ambient appearance of the environment [76].

They also proposed a method of superimposing a food onto the VE using chroma key composition (Figure 2.17) [20]. Multiple elderly users who ate together in the system felt that food quality was higher when eating with the VE than when eating alone, and that lifted their spirits. Perez *et al.* proposed a method of eating in the VE by coloring desks and eating utensils red, and performing color detection (Figure 2.18) [19]. However, such an approach requires changing the colors of the desk and food for color detection and chroma key composition in advance. Therefore, the system should be usable on an ordinary desk. In both studies, users of the metaverse were unable to use these systems because they could not be used on top of existing VR applications such as VRChat [†] and Mozilla Hubs [‡]. It should run simultaneously with existing VR applications. In Chap. 5, we developed food overlay system which detects and tracks the food region in the VST images. This system supports can be used with the OpenVR API, widely used in existing VR applications.

[†] VRChat, `https://hello.vrchat.com/`, last accessed March 17, 2023.

[‡] Mozilla Hubs, `https://hubs.mozilla.com/`, last accessed March 17, 2023.

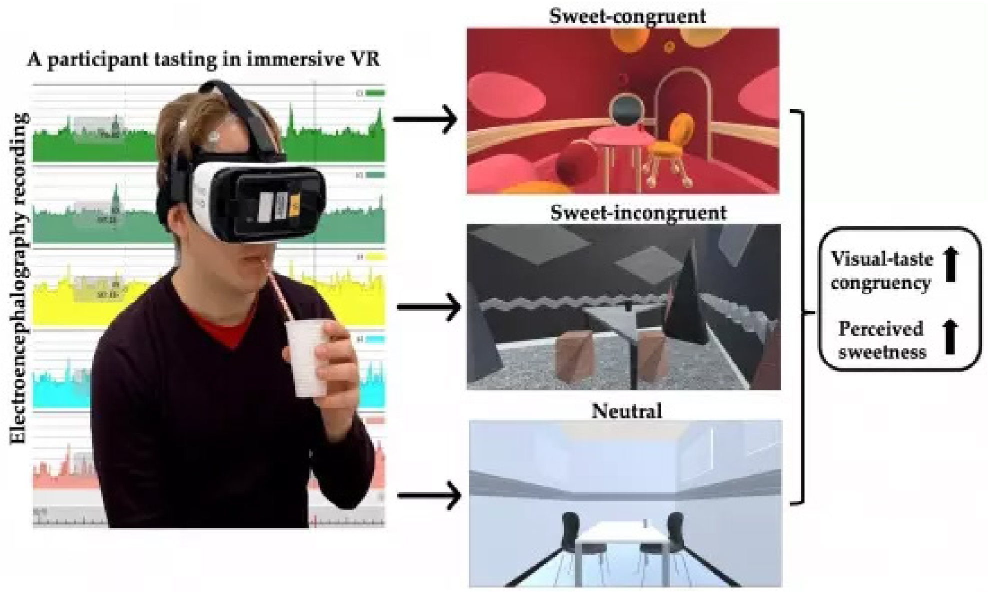

A participant tasting in immersive VR
Electroencephalography recording
Sweet-congruent
Sweet-incongruent
Neutral
Visual-taste congruency ↑
Perceived sweetness ↑

Figure 2.15: Rounded surroundings environment make sweet foods taste even sweeter [17].

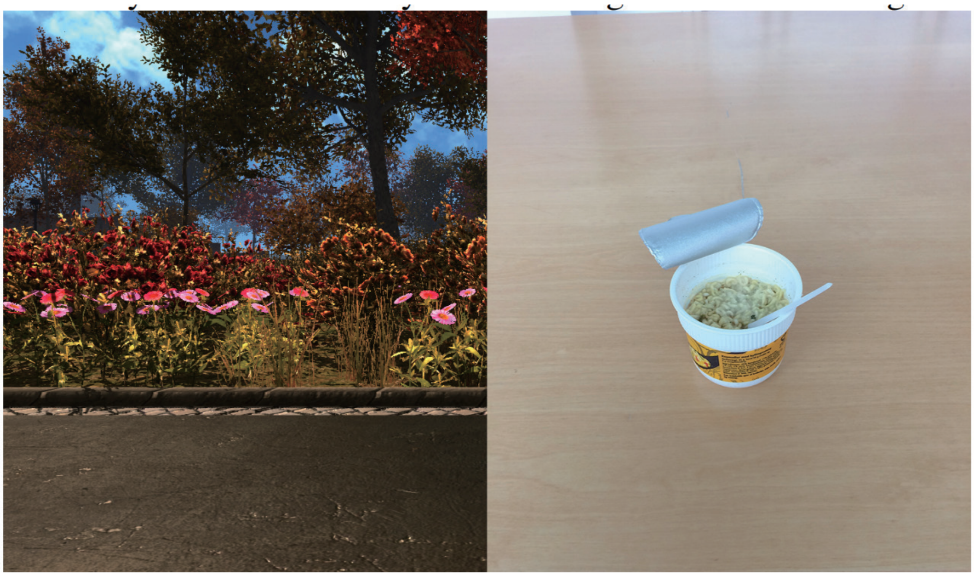


Figure 2.16: The transparency switches between the VE and VST by switching the transparency depending on the angle of the head. Left) The virtual park environment. Right) The visible image from the mounted camera [18].

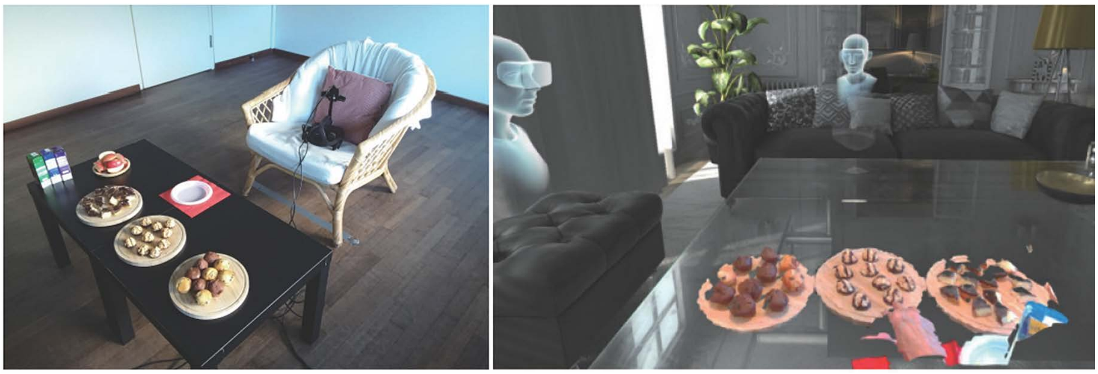


Figure 2.17: Eating together experience in the VE. Left) realistic eating environment. Right) food superimposed on the VE and the user's hand [20].

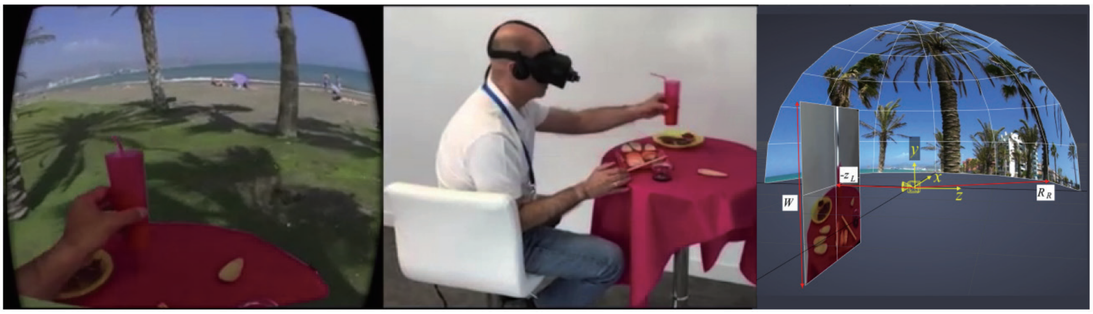


Figure 2.18: The user eats food with a red desk superimposed on the VE using red chrominance keying. Left) First-person view of the experience. Middle) External view of the experience. Right) Graphical model of the elements of the distributed reality [19].

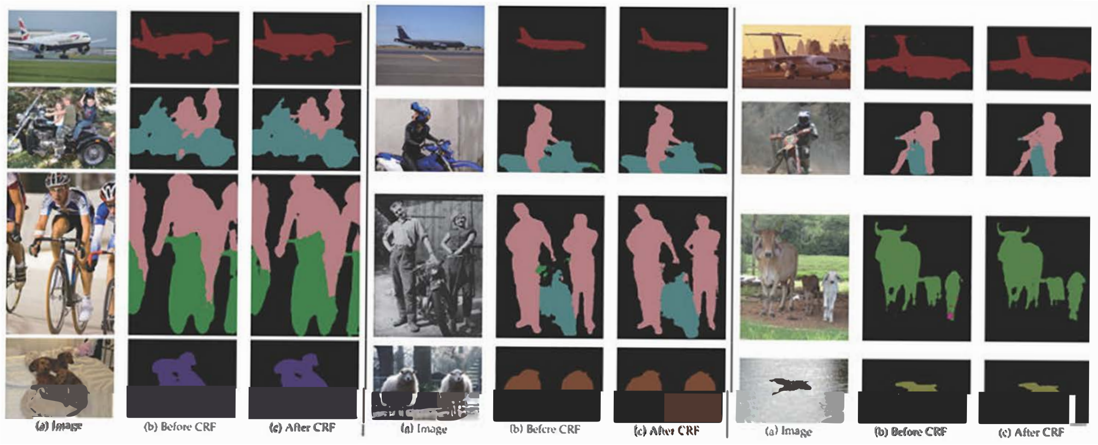


Figure 2.19: DeepLab results before and after input images and Conditional Random Fields (CRFs) when using the PASCAL VOC 2012 dataset [81].

2.5 Food Segmentation

We need to detect and segment only the food region from the image acquired from reality to superimpose only the food region on the VE. Semantic segmentation [80, 81] is the task of classifying pixels into semantic classes (e.g., humans, horses, etc.) within images (Figure 2.19). To accurately perform the task for food images, UEC-FoodPix [82, 83] contains paired 10,000 food images and fully annotated semantic masks (Figure 2.20). However, a neural network trained with UEC-FoodPix cannot be applied to real-time applications because an image dataset has no temporal information; its inconsistent outputs might result in discomfort.

To avoid this serious problem, Video Object Segmentation (VOS) [84, 85] performs to estimate a target location while keeping object correspondences between frames in a video. Specifically, it is given the arbitrary object location in the first frame of a video and predicts its position in all subsequent frames. SiamMask [85] is the first real-time and high-accurate VOS method. SiamMask can apply arbitrary objects without fine-tuning as long as it is initialized using a single bounding box. In Chap. 5, we use siamMask to track food regions to obtain food segmentation images.

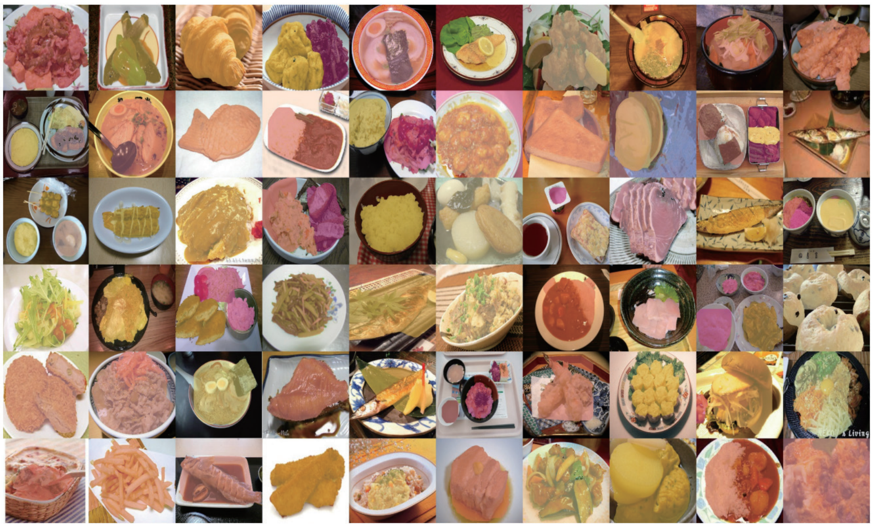


Figure 2.20: UECFood-100 images overlaid with segmentation masks annotated in UECFoodPix Complete [83].

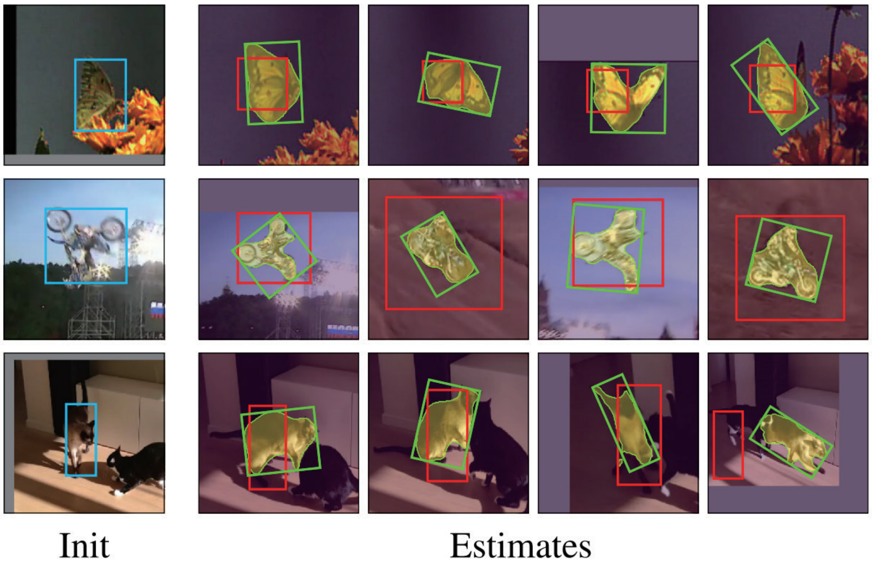


Figure 2.21: The result of SiamMask's tracking. Green frame) specifies a bounding box. Blue frame) around the target to be tracked in the first frame and automatically performs segmentation and tracking in the subsequent frames [85]. The frame box shows the tracking results of the Efficient Convolution Operators method [86].

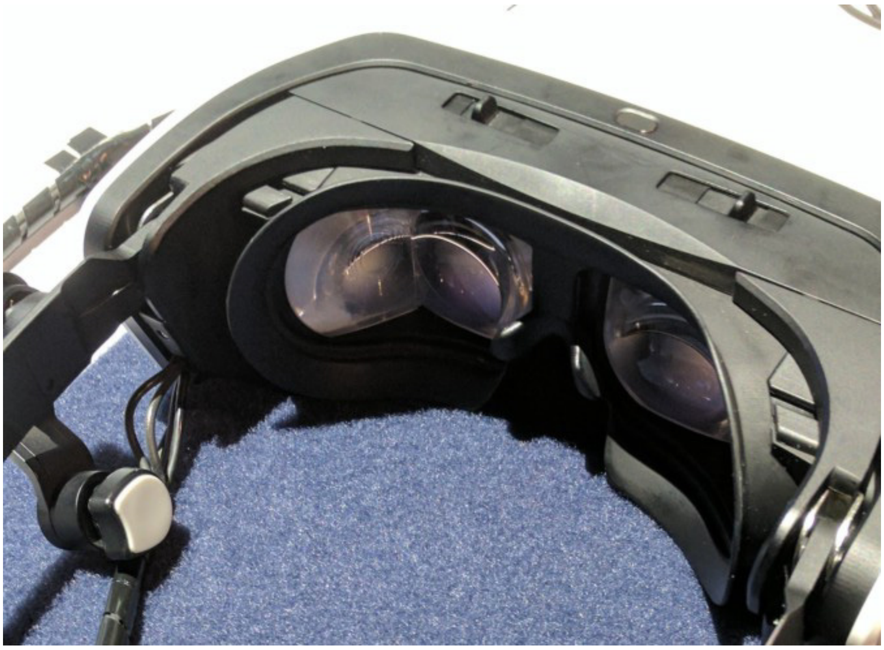


Figure 2.22: Panasonic HMD uses two fused lenses.

2.6 HMDs with a wide-FoV

We use HMDs to alter the appearance of food and the VE visually In particular, the downward FoV is necessary for the visual presentation of the near-body space close to our bodies and is essential for the visual presentation near the mouth and display of the avatar body for the multisensory flavor perception that we are studying. In addition, HMDs with a wide-FoV provide a high sense of presence, although the risk of cybersickness increases [87]. Therefore, research and development of wide-FoV HMDs such as StarVR One [§] and Pimax 8K Plus [¶] is actively pursued. Ratcliff *et al.* proposed a wide-FoV HMD using a curved display and a curved microlens array [88]. Panasonic has developed an HMD with a wide horizontal FoV using two displays per eye and two fused lenses (Figure 2.22) [‖]. Rakkolainen *et al.* developed a wide-FoV HMD using additional lenses and displays to the horizontal periphery (Figure 2.23) [89]. However, conventional wide-FoV HMDs are specifically for presenting frontal and horizontal images, making it difficult to present images in the downward FoV.

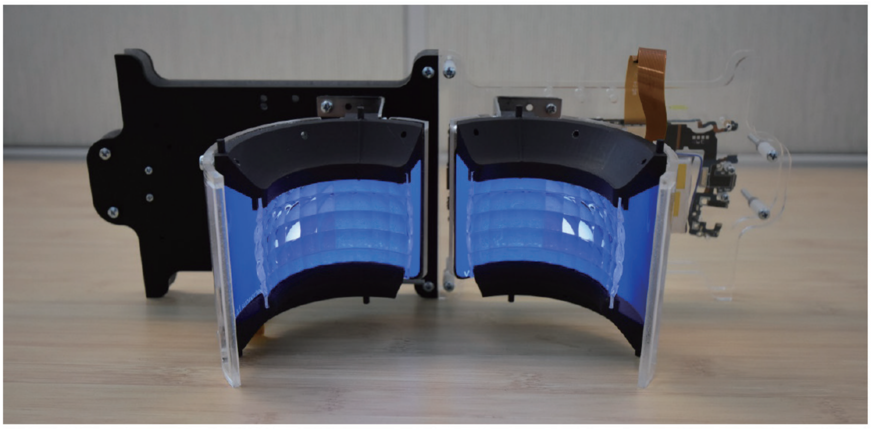


Figure 2.23: HMD with microlensarray for wide horizontal FoV [89].

In some studies, pseudo-FoV expansion methods have been proposed [90, 91] on the basis of the fact that the peripheral visual field is less receptive to information than the central visual field [92]. Xiao *et al.* placed a large number of full-color Liquid Crystal Displays (LCDs) and diffuser plates outside the eyepiece optics of the HMD and synchronized them with the HMD's image to expand the viewing angle (Figure 2.24) [90]. Yamada *et al.* combined a convex lens with normal magnification and a high magnification Fresnel lens to project a blurred image through the Fresnel lens in the peripheral FoV to expand the viewing angle inexpensively [91]. A downward FoV expansion is possible with this approach at the expense of low image resolution in the peripheral visual field.

Lindeman *et al.* developed a system that presents information about the RE in the downward view angle by installing an aperture under the eyepiece lens using an LC shutter with adjustable transparency [93, 94] for typing a real keyboard in VR, for example (Figure 2.25). The RE can be seen in the lower FoV with their system, but the virtual image cannot be presented in the aperture. Endo *et al.* developed an HMD at the same time as our study that increased the horizontal and vertical FoV by adding smartphones to the left, right, and bottom of the

[§] StarVR, `https://www.starvr.com/`, last accessed March 17, 2023.

[¶] Pimax, `https://www.pimaxvr.com/`, last accessed March 17, 2023.

[‖] Panasonic, `https://channel.panasonic.com/contents/19737/`, last accessed March 17, 2023.

[‖] MoguraVR, `https://www.moguravr.com/panasonic-vrhmd/`, last accessed March 17, 2023.

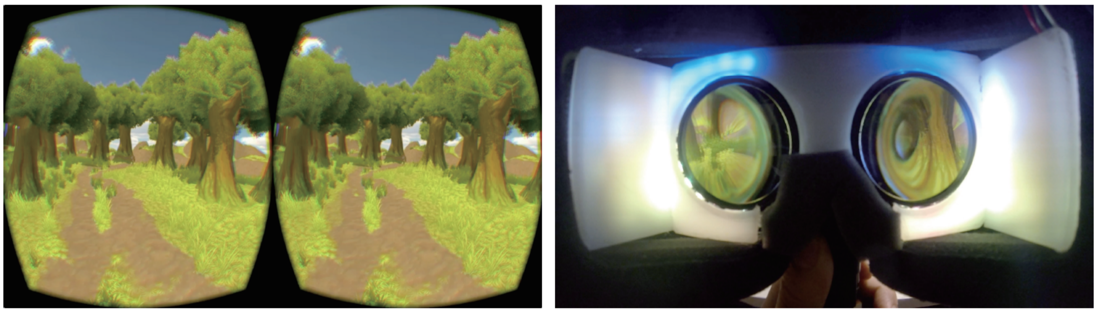


Figure 2.24: The forest scene is presented using an HMD that presents images in the peripheral vision. Left) Virtual scene rendered for Oculus Rift. Right) Scene displayed with the function to present images in the peripheral vision enabled [90].

HMD (Figure 2.26) [95]. It isn't easy to eat with this HMD because the added smartphone covers the mouth area completely.

As described above, most research on increasing the viewing angle of HMDs has focused on increasing the horizontal FoV, with little work done on increasing the vertical FoV. In particular, the downward FoV has not been considered important. Restriction of the downward FoV inhibits the presentation of visual information near the mouth, which is necessary for gustatory manipulation using visual modulation.

In Chap. 6, we develop an HMD with an increased downward FoV by adding two sets of lenses and displays in the vertically downward direction to an existing HMD to solve this problem. We also report on our investigation of the effects of presenting visual information near the mouth on multisensory flavor perception using an HMD with an increased downward FoV.

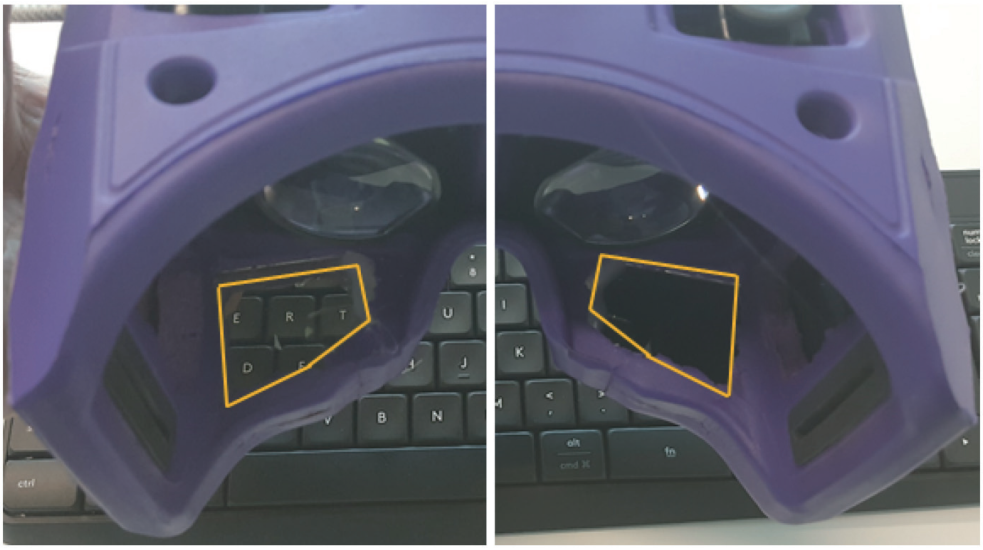


Figure 2.25: HMD with LC shutter installed at the bottom with changeable transparency, a keyboard of the real world can be displayed [94].

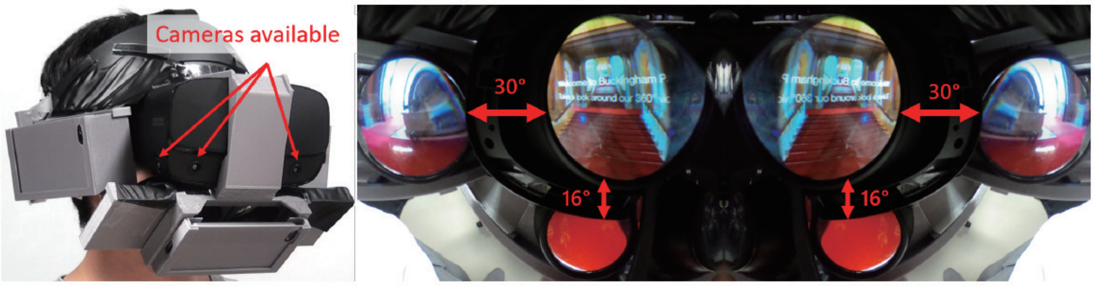


Figure 2.26: HMD with increased FoV by adding smartphones to the left, right, and bottom [95].

2.7 Information presented on the downward FoV

In this doctoral dissertation, which investigates the influence of gustatory manipulation by visual modulation on multisensory flavor perception, it is necessary to investigate the effect of body displays and appearance changes that are always visible in daily life. It is because the likelihood of information is important for humans to perform multisensory integration [21, 22], and the display of one's own body is essential to improve the likelihood of visual information. In addition, the appearance of the avatar used by the user changes the behavior and psychological state of the human being. For example, participants in the experiment use an avatar that looks like Einstein, their test scores improve (Figure 2.27) [31], and they beat the drums rhythmically when they use a casual dark-skin avatar that looks like a musician (Figure 2.28) [30]. Therefore, we plan to investigate how the visual appearance of avatars affects multisensory flavor perception. However, there is a critical problem with existing HMD-based avatar research. The problem is that the limited downward FoV makes it difficult to display bodies other than the avatar's hands.

The limited downward FoV of conventional HMDs degrades the SoE (SoSL, SoA, and SoBO) [23]. SoBO represents the perception that a virtual body is one's own through a visual self-avatar in the VE [96]. SoBO is also affected by the visual fidelity of the self-avatar. Basically, the closer the appearance of the self-avater is to one's own body, the higher the SoBO is evoked [23, 98]. For example, if one's own avatar's arms are human-like rather than robotic or spherical, it evokes a stronger SoBO (Figure 2.29) [97, 99, 100, 101, 102].

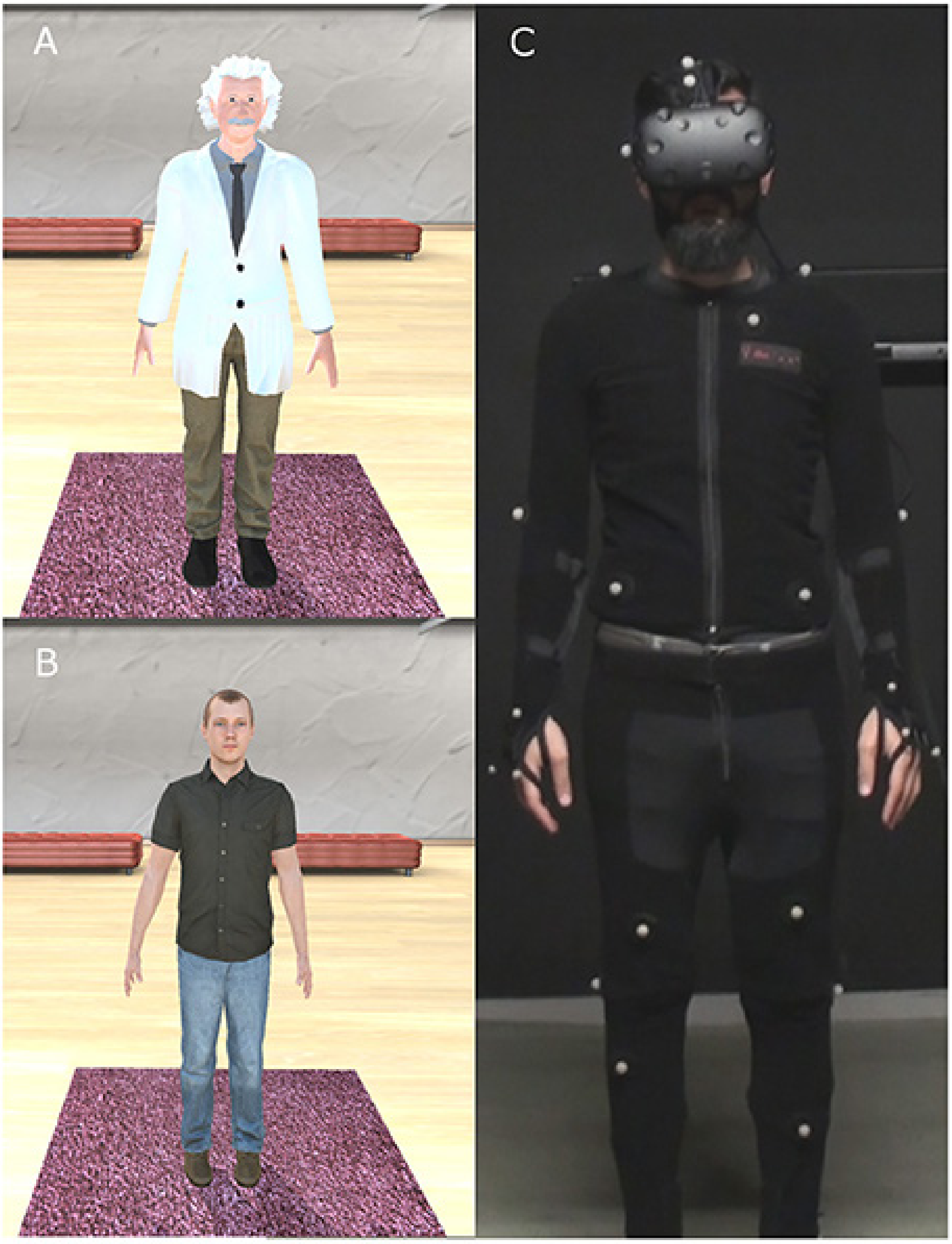


Figure 2.27: Participants who were embodied in a virtual body that signifies super-intelligence. Einstein) improved their test scores. A) Einstein's virtual body. B) Normal virtual body. C) Participants wore an HMD, and their body movements were tracked [31].

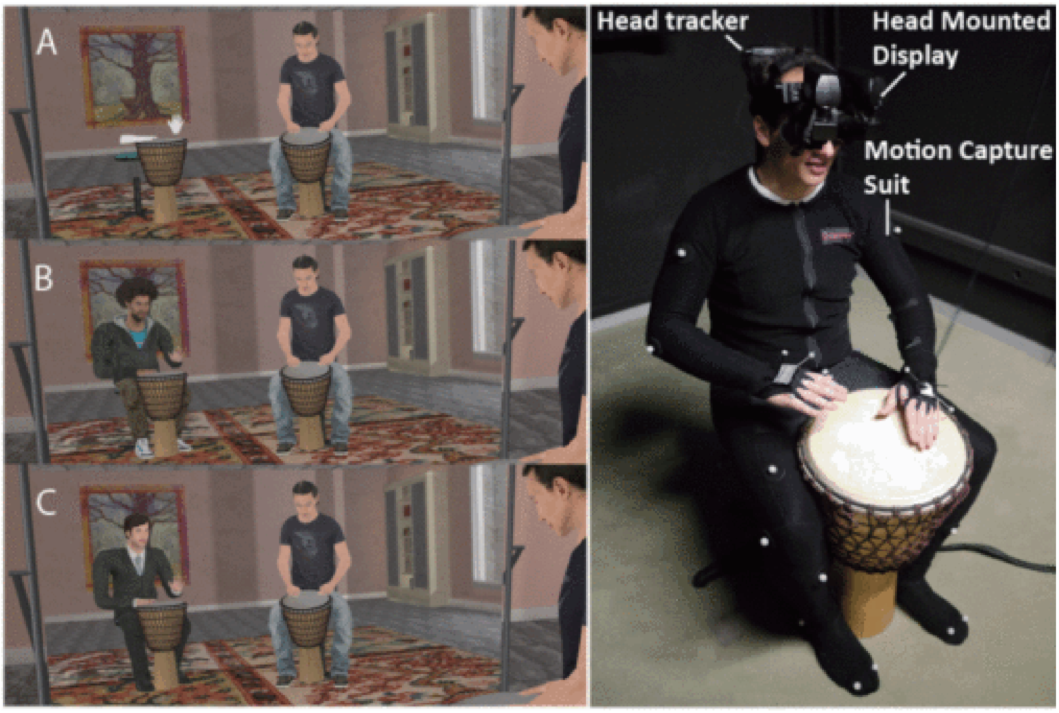


Figure 2.28: Participants with casual dark-skin avatars beat drums more rhythmically than in the other conditions. Left) Experimental conditions reflected in the mirror. A) White hand avatar. B) Casual dark-skinned avatar. C) Formal light-skinned avatar. Right) A participant wearing an HMD and a body-tracking suit, sitting on a stool and playing the drums [30].

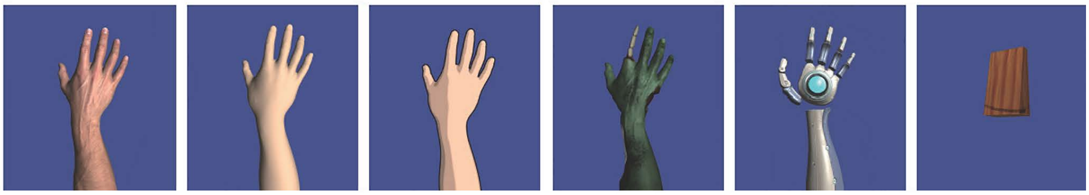


Figure 2.29: Participants strongly perceive virtual hand illusion [96] when using a realistic human hand. From left to right: realistic hand, toony hand, very toony hand, zombie hand, robot hand, wooden block [97].

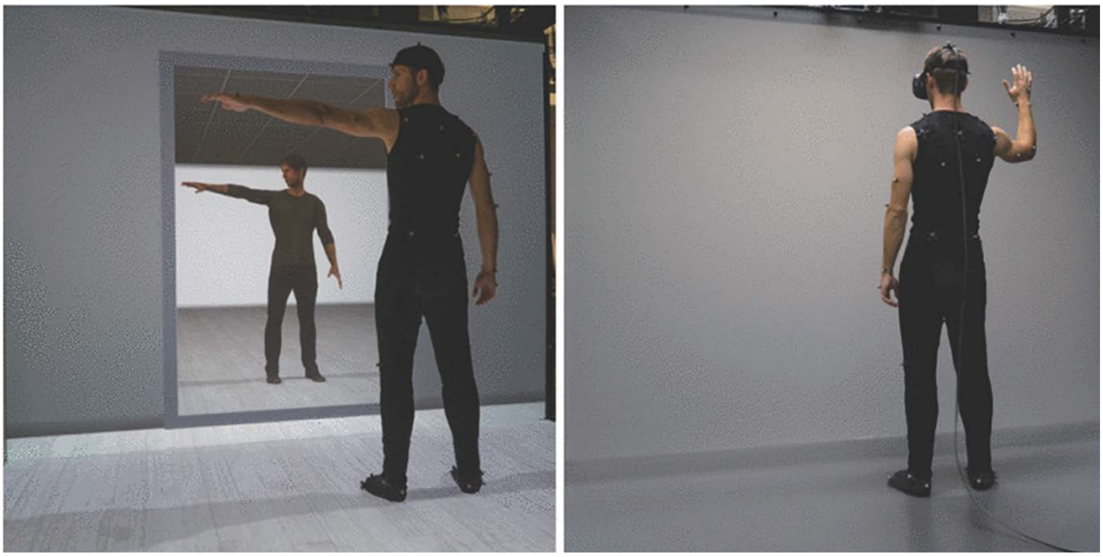


Figure 2.30: Full-body avatar is reflected in a virtual mirror. Left) Using CAVE. Right) Using HMD [103].

In experiments using full-body avatars, SoBO is often evoked using a virtual mirror in VEs (Figure 2.30) [24, 25, 26, 103]. McManus *et al.* reported that using a full-body humanoid avatar improved the performance of the object interaction task and the stepping stone task [27]. In their experiment, a virtual mirror always showed the view of the avatar. On the other hand, Streuber *et al.* reported that using a full-body humanoid avatar did not affect the performance in a task without the avatar shown via a virtual mirror [28]. They speculated that the difference in the results could be due to the fact that the participants could not grasp the position and posture of their self-avatars with the limited downward FoV [27].

In an experiment in which participants saw their self-avatars in a virtual mirror only before the task, Ogawa *et al.* reported that the participants using realistic full-body humanoid avatars more often avoided walking into virtual walls than those using low anthropomorphic avatars (hand or full-body robotic avatars), but there was no significant difference in perceived SoBO among avatar conditions [25]. Lugrin *et al.* also reported no significant difference in presence nor SoBO when comparing the use of three different types of self-avatars (controller, hands, and upper body) [104]. These studies suggest that although behavior changes under the influence of the initial impression provided by the virtual mirror, avatars must always be displayed to enhance the SoBO felt when acting in VEs. Therefore, expanding the downward FoV of the HMD is expected to be effective in inducing realistic behavior and enhancing the SoBO.

Hand–torso connectivity is considered to be one of the most important factors for evoking SoA and SoBO (Figure 2.31) [105, 106]. The limited downward FoV makes it difficult to display the shoulders and upper arms, which are important for the hand–torso connectivity. This could reduce the visibility of the hand–torso connectivity and decrease SoA and SoBO. Pan and Steed showed that visually displaying an animation of the self-avatar's feet following the real body improved presence, SoA, and SoBO, but the improvement was small [107]. One of the reasons for this small improvement is thought to be the inability to see the legs through conventional HMDs. Therefore, expanding the downward FoV with the visual presentation of the upper body, hand–torso connectivity, and feet of the self-avatar is expected to enhance SoA and SoBO.

Increasing the downward FoV could enhance not only SoA and SoBO but also SoSL because it improves visibility around the feet. Jones *et al.* found that increasing the vertical FoV improved the distance judgment in VR (Figure 2.32) [108]. Moreover, the presence or absence of a full-body avatar [109] and avatar visibility [110, 111] affect the ability to estimate the distance around oneself. On the other hand, Dewez *et al.* reported that in a task of tracing lines on the ground, the presence or absence of an avatar did not affect performance nor preference [112]. SoSL could be improved by making it easier to see where the avatar is in the VE by expanding the downward FoV.

A limited downward FoV also inhibits the user's behavior. For example, the limited downward FoV increases the downward head pitch angle when walking on rocks or slippery ground in the RE [113]. Even in VEs, the reduced FoV prolongs the time it takes to walk to the destination and increases the number of obstacle contacts [114]. In the RE, when descending stairs, a limited downward FoV decreases walking speed and increases downward head pitch angle [115]. On the other hand, in VEs, fear of falling may cause users to look at their feet more frequently than necessary when the downward FoV is limited. Considering these factors, it is expected that HMDs that provide an FoV close to the human FoV

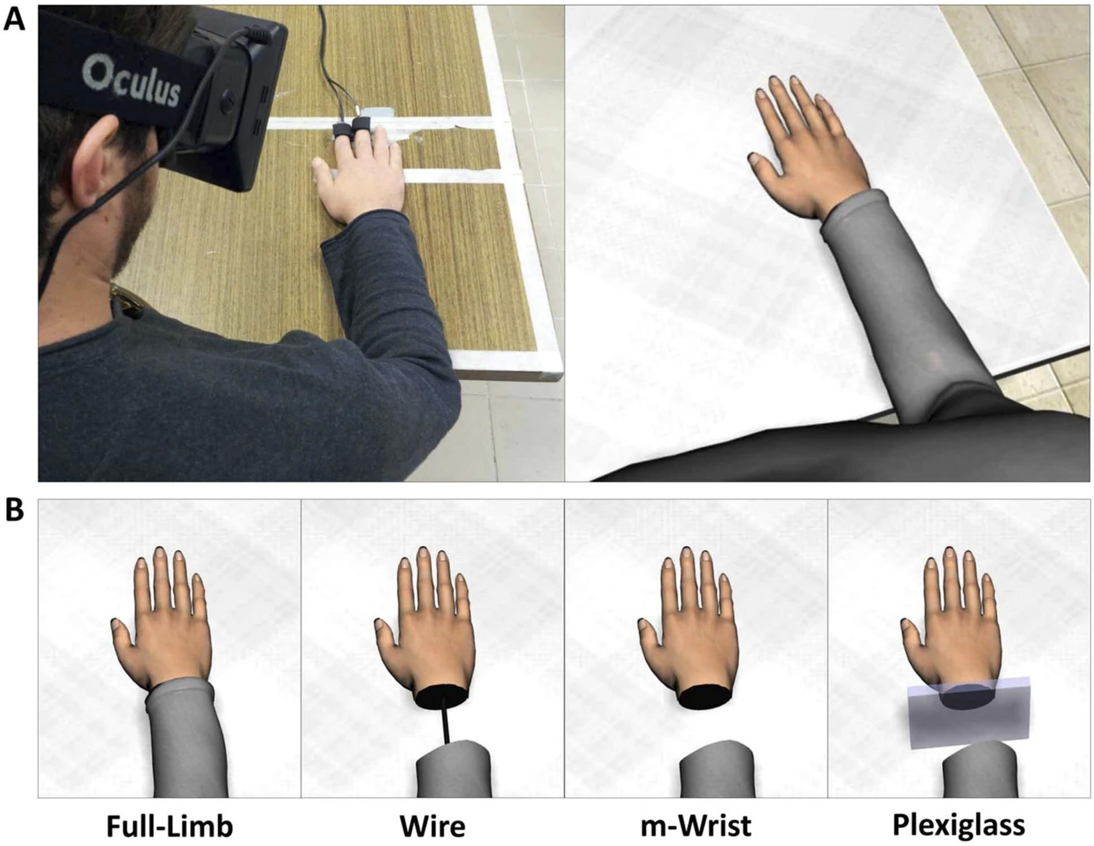


Figure 2.31: Changing hand and forearm connectivity revealed that only the natural fully connected virtual limb condition elicited high ownership. A) Participants observed in first-person perspective through an HMD the virtual body in the same location and posture as the physical one. B) The visual appearances of the avatar' s right limb, Full-Limb: A standard full arm, Wire: A limb with a thin black rigid wire connecting the forearm and the hand, m-Wrist: A limb with a missing wrist, Plexiglass: A limb with a missing wrist with a Plexiglass panel placed in the blank space between the hand and the forearm [106].

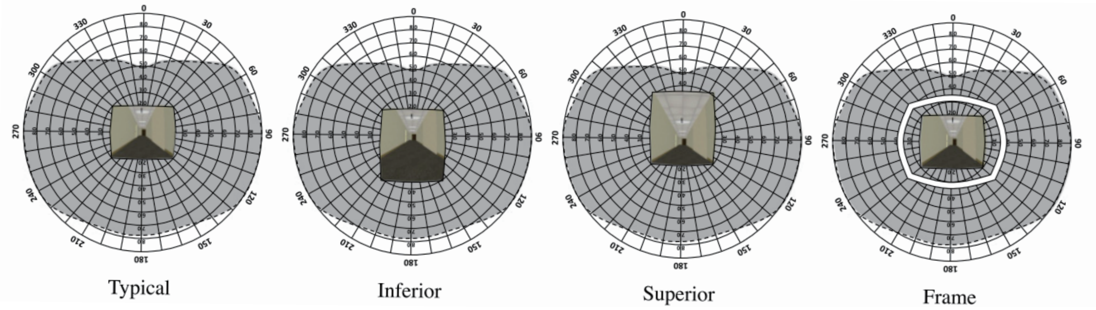


Figure 2.32: Participants' ability to judge distance is improved in the Inferior and Superior conditions with an increased vertical viewing angle. Typical) Normal vertical viewing angle. Inferior) Increased vertical viewing angle in the downward direction. Superior) Increased vertical viewing angle in the upward direction. Frame) White frame added outside the viewing angle, which is effective in improving the ability to judge distance [108].

in the real world will bring behavior closer to reality. On the other hand, in VEs, people behave differently in avoidance than in the RE. Sanz *et al.* found that users tend to avoid virtual objects faster than when avoiding real objects in a large immersive projection environment [116]. Pan and Steed found that when the self-avatar's feet did not follow the user's movements in VEs, the user moved their feet significantly further away from the obstacle than when they did [107]. Increasing the HMD's downward FoV makes it easier to notice discrepancies between the body and the environment, which may lead to the observation of behaviors that are different from those in reality and a reduced SoE.

As described above, increasing the downward FoV may affect presence and various aspects such as SoE (SoSL, SoA, and SoBO), head movement, and distance traveled. In Chap. 7, we experimentally investigated them comprehensively. In addition, we also investigated the effect of the increased downward FoV on cybersickness, because cybersickness tends to increase with an increasing FoV [90]. We believe that investigating the effectiveness of avatar displays at downward FoV and improving the usability and effectiveness of avatars will further facilitate research investigating multisensory flavor perception.

3 Manipulation of taste and type of food by changing its appearance

3.1 GAN-based Real-time Food-to-Food Translation and Its Impact on Vision-induced Gustatory Manipulation

In this chapter, we first report in detail a user study on an AR system that overlays a 3D model of a target food over a different type of original food and its successful experimental results. We then summarize the identified problems of 3D model-based systems and describe in detail the implementation of and the experiment resurt on the more flexible GAN-based AR system, which changes the appearance of one type of food into another in real time. The GAN has recently attracted considerable attention for the quality of its image-to-image translation. However, little is known as to whether it can be used to manipulate gustatory sensations and to what extent.

In this chapter, we show, for the first time, the great potential of GAN-based cross-modal effects as a simple yet powerful tool for multimodal AR systems. The major contributions of this chapter are as follows:

- We present a user study on a vision-induced gustatory manipulation system using a 3D food model and report its successful experimental results. We also identify its problems and discuss the necessity of a more flexible gustatory manipulation approach.

- We then present a more flexible vision-induced gustatory manipulation system using GAN-based real-time image-to-image translation. Our system flexibly supports multiple types of target food, which changes its appearance dynamically and interactively in accordance with the deformation of the original food.
- We finally report the experiment resurt on GAN-based gustatory manipulation with several combinations of food-to-food translation. Our experimental results confirmed that participants more strongly perceived the taste of the target food solely upon GAN-based visual modulation in addition to a few interesting episodes of cross-modal effects such as those between olfactory and tactile sensations.

3.2 Effectiveness and Problems of 3D Model-based System

In this section, we investigate whether the participants taste the type of food that they are seeing rather than the actual food that they are eating, using a 3D model-based system. Assuming that a more realistic appearance is more effective, we use a 3D reconstructed model of real food and overlay it onto a different type of food by means of VST AR.

3.2.1 Food Selection

As the target (destination) type of food, we chose *r*amen noodles for its popularity, specifically among East Asian countries. In addition, it is relatively common that a person cannot eat ramen noodles for nutritional (e.g., salt, fat, and lye water) and religious (e.g., pork) reasons. For example, a patient with Crohn's disease cannot consume lye water and thus cannot eat ramen noodles. Therefore, it is a good target for virtually making it possible to eat. We chose plain *s*omen noodles without fat or pork as an original (source) type of food as is also popular, is also made of wheat flour and has similar thickness to *r*amen noodles, and does not typically contain such 'forbidden' ingredients.

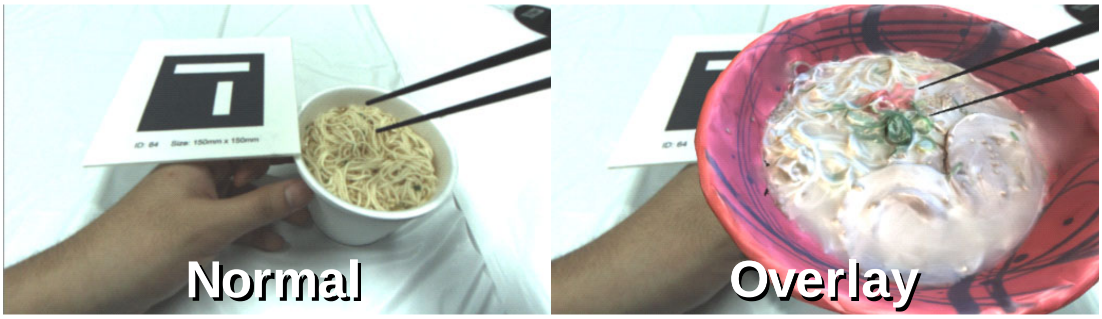


Figure 3.1: Left) Somen noodles in a cup with an AR marker (*N*ormal condition in the experiment). Right) Overlaid 3D model of ramen noodles (*O*verlay condition in the experiment). Although realistic, the model appearance does not change in accordance with the deformation of the actual food. It also causes an occlusion problem.

More specifically, we chose *t*onkotsu ramen noodles (known for their thick soup broth based on pork bones) available at university cafeterias, and instant somen noodles (without toppings) in a warm dashi broth (Nisshin-no-donbei from Nissin Inc.). Somen noodles served with warm soup is sometimes called nyumen and common in many regions in Japan even though it is slightly less common than those with cold soup. We used the name "somen noodles" in this experiment because the product we used is named so. It is also defined as "somen noodles" in the "Quality Labeling Standards for Dried Noodles [**]" established by the Consumer Affairs Agency of Japan.

3.2.2 System Overview

Our AR system for the first user study consists of a HMD (Oculus Rift CV1), a pair of stereo cameras (Ovrvision Pro, 1280×720 pixels at 60 fps), a desktop computer (Intel Core i7-4790K, 4.00GHz, 16GB, NVIDIA GTX1060), and a tin cup with an AR marker. The Unity-based software overlays a 3D model of ramen noodles over the cup. The 3D model was created with Autodesk ReMake and

[**] Quality Labeling Standards for Dried Noodles, `https://www.cao.go.jp/consumer/history/01/kabusoshiki/syokuhinhyouji/doc/004_101004_shiryou2-5.pdf`, last accessed March 17, 2023.

several photographs of the target tonkotsu ramen noodles (see Fig. 3.1(right)). The 3D model stays in the same position in space when the AR marker is lost, with the help of the HMD head-tracking system. For ease of eating, the black chopsticks are always made to appear in front of the 3D model by intensity thresholding. Olfactory displays are not used because we are interested in how much the gustation can be manipulated solely by visual modulation.

3.2.3 Procedure

It is common to investigate the effects of gustatory manipulation without showing the original food to the participants. However, we intentionally present the original food to each participant without visual modulation at the beginning. This is because one of our objective is to increase the QoL of people who are unable to eat what they want, and it is a natural assumption that they are aware of the original food (food substitute) they are actually eating. Note that this is a more challenging situation and it may have negative (smaller) effects on the extent of gustatory manipulation.

Each of the participants then puts on the HMD and touches the empty cup for three minutes to acclimate to the VST experience. Then they take two bites of somen noodles in one of two conditions, with (referred to as Rn) or without (referred to as Sn) the overlay of the 3D ramen model. They are orally informed which of the two conditions (Sn or Rn) they are experiencing. They then drink some water to clean the mouth and take two bites in the other condition. The order of the two conditions is randomized. They wear the HMD even in the Sn condition to minimize unwanted differences between the two conditions and highlight the impact of the visual modulation.

Finally, they answer a questionnaire that consists of the following five questions on a visual analog scale (VAS) (0 and 100 being 'strongly disagree' and 'strongly agree', respectively). A VAS is commonly used to measure the intensity of a person's gustation [117]. Participants answer **Q1** to **Q4** after each condition and **Q5** after all the conditions.

Q1. It tasted like $s$omen noodles in the Sn condition.

Q2. It tasted like $r$amen noodles in the Sn condition.

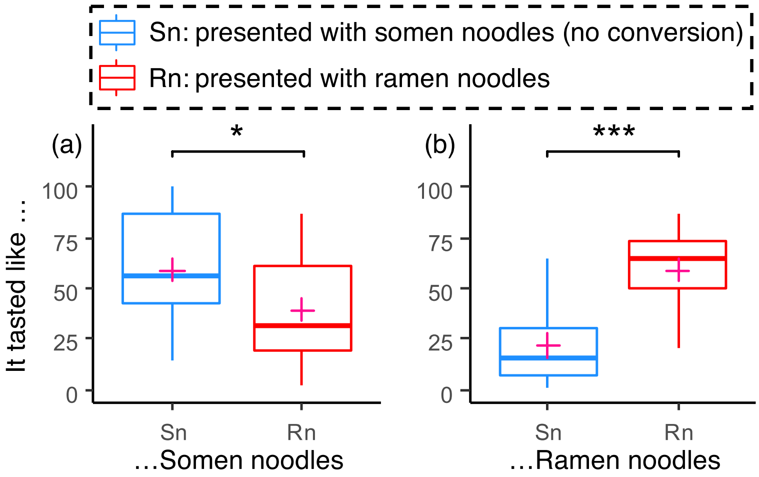


Figure 3.2: Results of the preliminary experiment.

Q3. It tasted like *s*omen noodles in the *R*n condition.

Q4. It tasted like *r*amen noodles in the *R*n condition.

Q5. It was easy to eat noodles with the HMD.

Eighteen volunteers (16 males and two females) ranging in age from 18 to 39 participated in the experiment. The purpose and the procedure were orally explained to and agreed by each participant. The experiment was approved by the ethics committee.The participants were 12 Japanese, four French, and two Thai people recruited from our university. All participants had eaten and were familiar with both somen and ramen noodles before taking part in the experiment.The results for **Q1** and **Q3** on the strength of the perceived taste of somen noodles are $M = 58.8, SD = 26.5$ and $M = 39.1, SD = 26.6$, respectively (see Fig. 3.2 left). Regarding the cultural differences, the results for **Q1** and **Q3** are $M = 48.8, SD = 24.5$ and $M = 25.6, SD = 18.3$ for Japanese, and $M = 79.0, SD = 17.2$ and $M = 66.0, SD = 19.1$ for non-Japanese, respectively. A two-way ANOVA revealed that there are statistically significant differences in both the food types ($F(1, 32) = 7.415, p < 0.05$) and food culture conditions ($F(1, 32) = 21.036, p < 0.001$). However, no significant interaction was found.

The results for **Q2** and **Q4** on the strength of the perceived taste of ramen noodles are $M = 21.6, SD = 18.5$ and $M = 58.9, SD = 18.1$, respectively (see

Fig. 3.2 right). Regarding the cultural differences, the results for **Q2** and **Q4** are $M = 24.8, SD = 20.9$ and $M = 64.7, SD = 12.8$ for Japanese, and $M = 15.3, SD = 9.74$ and $M = 47.3, SD = 21.4$ for non-Japanese, respectively. A two-way ANOVA revealed that there are statistically significant differences (statistically significant differences) in both the food types ($F(1, 32) = 38.001, p < 0.001$) and food culture conditions ($F(1, 32) = 4.348, p < 0.05$). However, no significant interaction was found.

The result for **Q5** is $M = 27.6, SD = 19.4$ and most of the participants found it difficult to eat noodles with the HMD.

3.2.4 Discussion and Identified Problems

The experimental results clearly show that the participants felt the taste of the target food more strongly and the taste of the original food less strongry.Therefore, the effectiveness of the vision-induced gustatory manipulation system between different types of food (noodles) is confirmed.

The strengths of the taste perceived by Japanese participants were different from those perceived by non-Japanese participants. We believe this is due to the differences in memories evoked by the differences in their food cultures. It has been suggested that the perceived taste varies with the nationality [60, 61]. However, these results also indicate that vision-induced gustatory manipulation is effective regardless of the participants' food culture.

In a post hoc interview, each of the participants was additionally asked if they felt any change in taste between Sn and Rn. Three out of 18 participants did not feel any change in taste when presented with a 3D ramen model. However, two of these three participants commented that they did feel like they were eating the type of food that they were seeing. One of them mentioned "I didn't feel the change in taste. When I was seeing somen noodles, I felt that the type of food was somen noodles. When I was seeing ramen noodles, however, I felt that the type of food I was eating was ramen noodles."Visual manipulation of food seems effective in changing food recognition (what you think you are eating) even when the taste perception does not change.

The results also suggest that it was very difficult to eat noodles with the HMD. The participants commented that it was particularly difficult with the 3D model

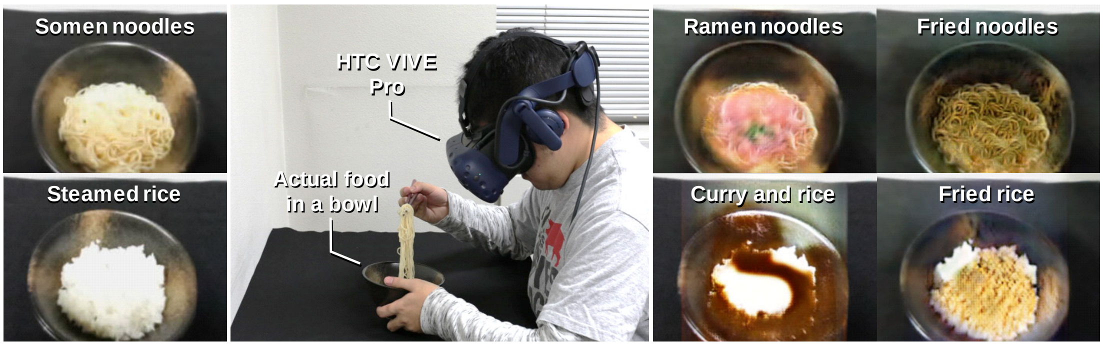


Figure 3.3: Our GAN-based real time food-to-food translation system in action. Left) Input food images. Middle) User with a VST HMD experiencing vision-induced gustatory manipulation. Right) Examples of translated images.

overlay, because the appearance of the 3D model did not change dynamically and interactively in accordance with the deformation of the food that they were actually eating and because the 3D model occluded the food as well as the user's hands. Another limitation of the system is that overlaying the same single 3D food model is inflexible and boring. A 3D model needs to be prepared for each food type in advance of usage.

3.3 GAN-based Real-time Food-to-food Translation System

In this section, we will describe the implementation details of the GAN-based real-time food-to-food translation system we developed [118]. Our system addresses the problems encountered in the first user study and has the following benefits (see also Figs. 3.3 and 3.8 as well as the supplementary video). These characteristics will contribute to better flexibility and applicability, as well as ease of eating.

- The original food and the user's hand are not occluded unlike the case of the 3D model-based system, because the input video is modulated while retaining its visual features to some extent.

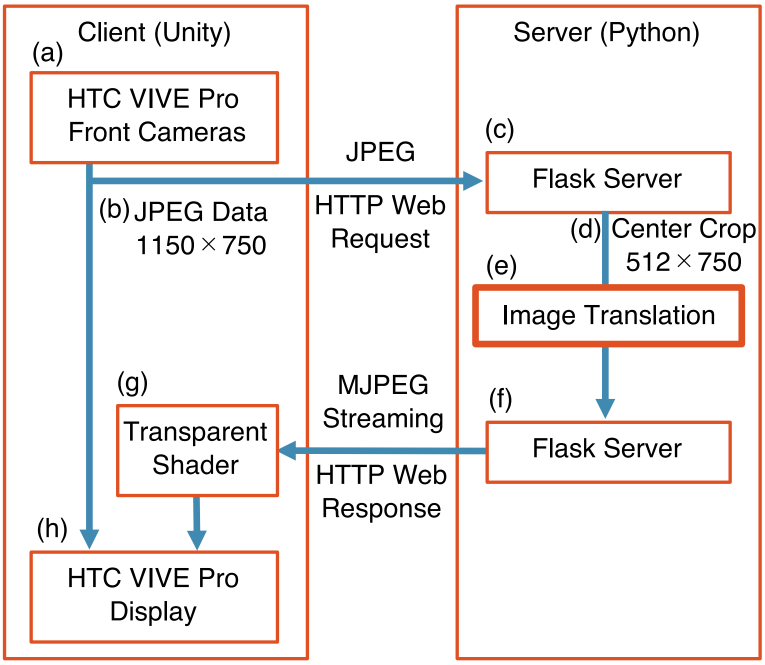


Figure 3.4: System configuration of the GAN-based food-to-food translation system. The client module acquires an RGB image from the camera (a), sends it to the server (b), overlays the processed image on the video background (g), and presents the scene to the user (h). The server module receives the sent image (c), crops the center (d), translates it to another food image (e), and sends it back (f).

- The appearance of the target food is dynamic and interactive in accordance with the deformation of the original food.
- Trained with multiple domains, the single GAN can support multiple combinations of food-to-food translation at the same time.

3.3.1 System Overview

Our GAN-based food-to-food translation system consists of client and server modules (see Fig. 3.4). They run on different desktop computers to maximize the overall performance. The client module is responsible for the front-end of the user interaction. It first acquires an RGB image from a front camera of the VST

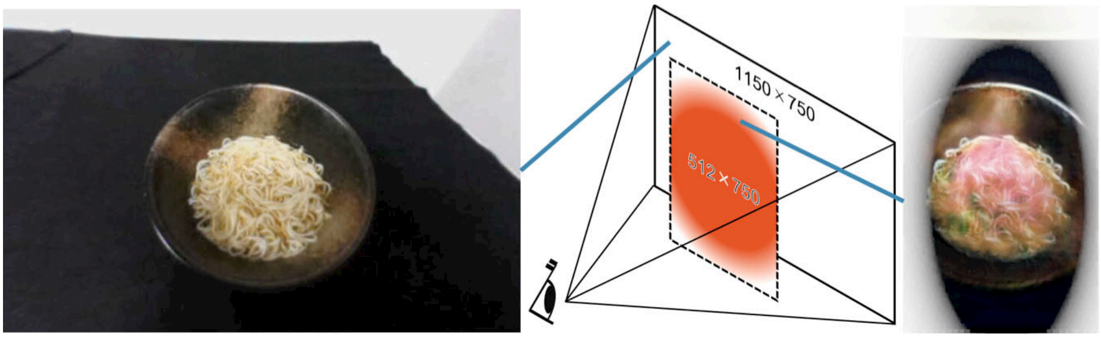


Figure 3.5: Left) Video background. Middle) schematic view of the image overlay. Right) corresponding translated image with radial gradient transparency.

HMD (Fig. 3.4(a)), then sends it to the server module (Fig. 3.4(b)), overlays the processed image over the stereo video background (Fig. 3.4(g)), and presents the scene to the user via the HMD (Fig. 3.4(h)).

The server module is responsible for the back-end of the image conversion. It first receives an RGB image from the client (Fig. 3.4(c)), crops the center (Fig. 3.4(d)), translates it to another food image (Fig. 3.4(e)), and sends it back to the client module (Fig. 3.4(f)). The implementation details of the client and server modules follow in the next subsections.

3.3.2 Client Module

The client module is implemented with Unity, and it runs on a desktop computer (Intel Core i7-8700K, 3.70GHz, 16GB, NVIDIA GTX1080 × 2) connected to a VST HMD (HTC VIVE Pro). The left and right front cameras of the HMD capture a pair of stereo images (1150×750 each) at 30 fps. However, we found that this was too slow to process all of the images on the server side at the full frame rate. As a good compromise, we send only the left image in JPEG format to the server module at 6 fps.

The client module then receives the translated image (512×750) via an MJPEG streamer and overlays it as a Unity Quad object onto the center of the video background for the left and right eyes (see Fig. 3.5). We only convert the middle of the image to reduce the processing time on the server side, retain the peripheral

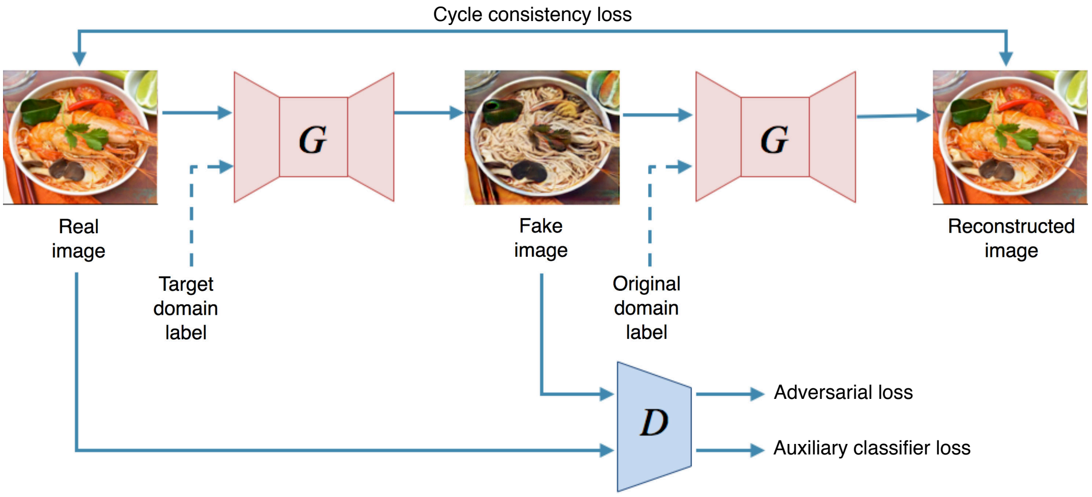


Figure 3.6: Overall architecture of StarGAN [73] model, consisting of two modules: a discriminator D and a generator G. The discriminator D learns to distinguish whether the input image is fake or real and classifies it into the class it belongs to. The generator G learns to generate images that deceive D.

view with minimal latency at 30 fps, and suppress motion sickness [119].

In spite of the relatively large latency of around 400 [ms] from the image capture to the overlay of the translated image, no participants appeared to have or reported simulator sickness throughout the experiment. The width of the translated image was sufficient to see the entire bowl and food. We also perform Shader-based radial gradient blending using the two-dimensional (2D) Gaussian distribution function in Eq.(3.1) with empirically tuned parameters of $\sigma = 0.3$, $x_0 = 0.03$, and $y_0 = 0.15$ for a comfortable visual experience.

$$f(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{1}{2}\left[\left(\frac{x-x_0}{\sigma}\right)^2 + \left(\frac{y-y_0}{\sigma}\right)^2\right]} \tag{3.1}$$

3.3.3 Server Module

The server module is implemented with Python, and it runs on the same computer as in the first user study. We translate food images in real time through the POST method of HTTP with the local web server. In the following, we explain the overall structure of the server module, then we introduce StarGAN, a framework that can transform multiple categories (see Fig. 3.6).

Preprocessing

The server module first receives a left front camera image (1150×750) of the HMD using Flask, a Python web application framework. Then it crops the center region (512×750) by a PyTorch function. We carry out the cropping on the server side because we find the overall performance to be faster than by doing it on Unity. We do not extract the food and bowl regions and translate the entire cropped image because precise extraction is difficult, particularly around moving chopsticks or a spoon. The translated image is streamed back to the client module in MJPEG format.

StarGAN

Model Objectives. Figure 3.6 shows the network of StarGAN [73]. The goal of the generator is to translate an input image x into a generated image $G(x, c)$, which is appropriately classified as the target domain c, using both the input image x and the target domain label c. In addition, the generator learns more than one domain translation at the same time. To achieve this, the discriminator performs not only the task of distinguishing real and fake images with an adversarial loss (Eq.(3.2)) but also the domain classification task of classifying the domain of the output image $G(x, c)$ of image x after translation by the generater with domain classification losses (Eqs.(3.3) and (3.4)) defined as

$$L_{adv} = \mathbb{E}_{x\sim\mathbb{P}_r}\left[\log D_{adv}(x)\right] + \mathbb{E}_{x,c\sim\mathbb{P}_r}[\log\left(1 - D_{adv}(G(x, c))\right)], \tag{3.2}$$

$$L_{cls}^{real} = \mathbb{E}_{x,c'\sim\mathbb{P}_r}\left[-\log D_{cls}(c'|x)\right], \tag{3.3}$$

$$L_{cls}^{fake} = \mathbb{E}_{x,c\sim\mathbb{P}_r}\left[-\log D_{cls}(c|G(x, c))\right]. \tag{3.4}$$

Here, G produces a fake image using both the input image x and the original domain label c, which are sampled from the data distribution $\mathbb{P}_r$, D_{adv} represents a discriminator that identifies real or fake, and D_{cls} represents a discriminator that identifies which class the translated image belongs to. $\mathbb{E}_x$ is the expected value and calculated as $\int_x f(x)dx$.

However, even if the generator attempts to translate the input image x into the output image $G(x, c)$ by minimizing only the objective functions (Eqs. (3.2) and (3.4)), it does not guarantee that the shape of the input image x will be maintained. To address this problem, a cycle consistency loss [67, 68] is applied to the objective function of the generator, defined as

$$L_{rec} = \mathbb{E}_{x,c,c' \sim \mathbb{P}_r}[\|x - G(G(x, c), c')\|_1], \tag{3.5}$$

where G reconstructs the original image $G(G(x, c), c')$ with the generated image $G(x, c)$ and the original domain label c'. Furthermore, to stabilize the learning and generate a higher quality image, the objective function using the Wasserstein distance with a gradient penalty [63, 120] is employed instead of Eq.(3.2), defined as

$$\begin{aligned} L_{adv} = \quad & \mathbb{E}_{x \sim \mathbb{P}_r}\left[D_{adv}(x)\right] - \mathbb{E}_{\hat{x} \sim \mathbb{P}_g}\left[D_{adv}(\hat{x})\right] - \\ & \lambda_{gp}\mathbb{E}_{\hat{x} \sim \mathbb{P}_{\hat{x}}}[(\|\nabla_{\hat{x}} D_{adv}(\hat{x})\|_2 - 1)^2], \end{aligned} \tag{3.6}$$

where $\hat{x}$ is sampled uniformly along straight lines between pairs of points sampled from the training data distribution $\mathbb{P}_r$ and the generator distribution $\mathbb{P}_g$. Finally, the objective functions of StarGAN are represented as

$$L_D = -L_{adv} + \lambda_{cls} L_{cls}^{real}, \tag{3.7}$$

$$L_G = L_{adv} + \lambda_{cls} L_{cls}^{fake} + \lambda_{rec} L_{rec}, \tag{3.8}$$

where λ_{cls} and λ_{rec} respectively control the relative importance of the classification and reconstruction objective functions. We use $\lambda_{cls} = 1$, $\lambda_{rec} = 10$, and $\lambda_{gp} = 10$ throughout the experiment. In this study, we learn the model on the basis of these losses.

Dataset. By adding a cycle consistency loss, the generator translates the image while preserving its shape. Thus, we have constrained the same structure ‘bowl’ so that a cycle consistency loss can be learned appropriately. We build a dataset of food images using UECFOOD-100 [74] and additional images we collected from the Twitter stream. However, the initial dataset included duplicate images and images in unnecessary domains. Therefore, to clean the dataset, we extracted the image features using VGG [121], clustered them using X-means, and removed unnecessary classes and duplicate images. The final dataset contains 149,370 images in five categories, as shown in Table 3.1. All the images in the dataset were resized to square RGB images of $256 \times 256 \times 3$, and their color values were normalized to $[-1, 1]$.

Network Architecture. The generator network has two convolution layers with a kernel size of 4, a stride size of 2, and a padding size of 1 for downsampling; six residual blocks [122], and two transposed convolution layers with a kernel size of 4, a stride size of 2, and a padding size of 1 for upsampling. We use instance normalization [123] followed by ReLU in all the convolutional layers. We use tanh as the activation function of the output layer. The discriminator network has five convolution layers with a kernel size of 4, a stride size of 2, and a padding size of 1. We use Leaky ReLU after all the convolutional layers. See [73] for the details of StarGAN.

Implementation Details. We use Adam as the optimizer of the generator and the discriminator with $\beta_1 = 0.5$, $\beta_2 = 0.999$. We use a learning rate of 0.0001, which is fixed throughout the training procedure. The batch size is set to 32.

3.4 Experiment

The purpose of the experiment is to investigate the impact of the GAN-based food-to-food translation system on vision-induced gustatory manipulation. We chose *s*teamed white rice as another original (source) food in addition to somen noodles because it is one of the most basic foods in the daily lives of East Asian people. In addition, it does not have a strong taste or flavor, and rice allergy is relatively rare. We investigate whether somen noodles can taste like ramen noodles or fried noodles (yakisoba) and whether steamed rice can taste like curry and rice or fried rice (chahan). Our StarGAN has learned all of these types of food except somen noodles. Somen noodle images were not included in the dataset. However, one of the advantageous properties of a GAN is that it can produce reasonable results from unknown input images.

For stable visual modulation, the experiment is conducted in a quiet room near a white wall, and the food is presented on a black table in a black bowl with red chopsticks or a spoon. Each of the participants is confirmed to be healthy and not too hungry or full. They first look at the original food in a bowl without the HMD. This is done for the same reason as in the first user study, that is, it is a natural assumption in our target scenario that the users are aware of the original food (food substitute) they are actually eating. Then they put on the HMD and look at the bowl with the presented food (either the original food or one of the two target types of food) for three minutes to acclimate to the system and the VST experience. Then, they are asked to state what food they are seeing *b*efore eating. We then tell them the correct answer and they answer whether or not this appears to be the case. Then they drink some water and take two bites of the food. They finally remove the HMD and answer the seven questions (**Q1** to **Q7**) below. They repeat the procedure three times each in a randomized order for steamed rice and somen noodles. The procedure is carried out on different days in a randomized order for the rice and noodle conditions to avoid unwanted confusion and interactions.

The questionnaire consists of the following seven questions on a VAS (0 and 100 being ‘strongly disagree’ and ‘strongly agree’, respectively). We added category

Table 3.1: Image dataset used to learn the network.

Category	# of images
Ramen noodles	75,350
Fried noodles	28,400
Steamed rice	7,390
Curry & rice	9,830
Fried rice	28,400
Total	149,370

questions (**Q4** to **Q6**) to ask what participants thought they were eating. This is because some participants felt that they were eating what they were seeing even when their perceived taste had not changed in the first user study. Participants answer **Q1** to **Q6** after each condition and **Q7** after all the conditions.

Q1. It tasted like *s*omen noodles (or *s*teamed rice).

Q2. It tasted like *r*amen noodles (or *c*urry and rice).

Q3. It tasted like *f*ried noodles (or *f*ried rice).

Q4. It felt like I was eating *s*omen noodles (or *s*teamed rice).

Q5. It felt like I was eating *r*amen noodles (or *c*urry and rice).

Q6. It felt like I was eating *f*ried noodles (or *f*ried rice).

Q7. It was easy to eat with the HMD.

3.5 Results and Discussion

3.5.1 Overview

All available Japanese participants in the first user study participated in the experiment for evaluating the effect of the greater ease of eating. Two of them were not available due to time constraints so two participants who had previously used the system, as described in Sec. 3, were additionally recruited. In the end, 12 volunteers (10 males and two females) ranging in age from 21 to 39 participated in the experiment. The purpose and the procedure were orally explained to and agreed by each participant. The experiment was approved by the ethics committee. Participants had previously eaten all of the six types of food used in the experiment. We excluded the foreign participants from the experiment because they had all never eaten at least one of the six types of food. In the following, the results and discussion are given for the noodle condisions followed by the rice conditions. We performed a one-way ANOVA followed by a post hoc analysis with the Holm–Bonferroni correction throughout the experiment.

3.5.2 Noodle Conditions

Results. Hereinafter, we denote the somen (original), ramen, and fried noodle conditions by the symbols Sn, Rn, and Fn, respectively. Figure 3.7 shows the results of these noodle conditions. The box plots in the upper and lower rows correspond to the results for **Q1** to **Q3** and **Q4** to **Q6**, respectively.

Fig. 3.7(a) shows the strength of the perceived taste of $s$omen noodles in the three visual modulation conditions Sn, Rn, and Fn. An ANOVA found a marginal trend toward significance ($F(2, 33) = 2.9328, p < 0.10$) and the Holm–Bonferroni method found a marginal trend toward significance between Sn and Fn ($p < 0.10$). Fig. 3.7(d) shows the confidence of the food being recognized as $s$omen noodles in the three visual modulation conditions Sn, Rn, and Fn. An ANOVA found an statistically significant differences ($F(2, 33) = 8.9314, p < 0.001$) and the Holm–Bonferroni method found statistically significant differences between Sn and Rn ($p < 0.01$) and between Sn and Fn ($p < 0.01$).

Fig. 3.7(b) shows the strength of the perceived taste of $r$amen noodles in Sn, Rn, and Fn. An ANOVA found an statistically significant differences ($F(2, 33) = 3.7006, p < 0.05$) and the Holm–Bonferroni method found a marginal trend toward significance and an statistically significant differences between Sn and Rn ($p < 0.10$) and between Rn and Fn ($p < 0.01$).

Fig. 3.7(e) shows the confidence of the recognized food as $r$amen noodles in Sn, Rn, and Fn. An ANOVA found an statistically significant differences ($F(2, 33) = 5.9197, p < 0.001$) and the Holm–Bonferroni method found statistically significant differences between Sn and Fn ($p < 0.05$) and between Rn and Fn ($p < 0.01$), as well as a marginal trend toward significance between Sn and Rn ($p < 0.10$).

Fig. 3.7(c) shows the strength of the perceived taste of $f$ried noodles in Sn, Rn, and Fn.

An ANOVA found an statistically significant differences ($F(2, 33) = 18.956, p < 0.001$) and the Holm–Bonferroni method found statistically significant differences between Sn and Fn ($p < 0.01$) and between Rn and Fn ($p < 0.01$).

Fig. 3.7(f) shows the confidence of the recognized food as $f$ried noodles in Sn, Rn, and Fn.

An ANOVA found an statistically significant differences ($F(2, 33) = 49.858, p < 0.001$) and the Holm–Bonferroni method found statistically significant differences

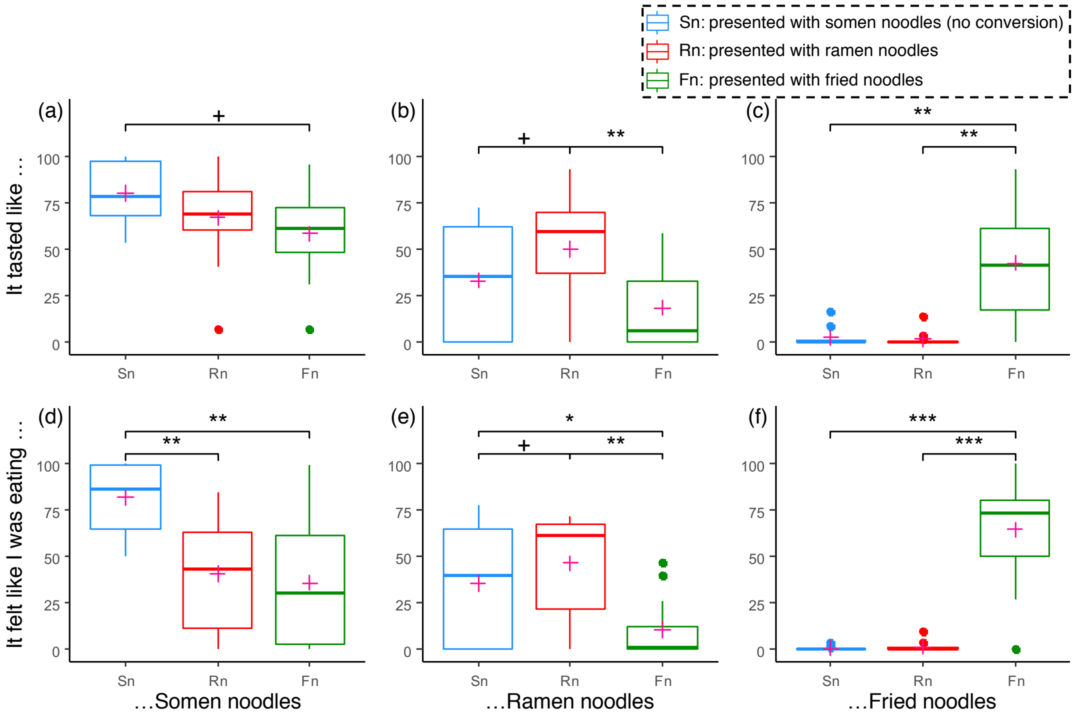


Figure 3.7: Results for various noodle conditions. The box plots in the upper and lower rows correspond to the results for **Q1** to **Q3** and **Q4** to **Q6**, respectively.

between Sn and Fn ($p < 0.001$) and between Rn and Fn ($p < 0.001$).

In the case of Rn, only six out of 12 participants (50%) thought that they were seeing ramen noodles at the beginning. In the case of Fn, all 12 participants (100%) thought that they were seeing fried noodles at the beginning. For Rn and Fn (and Cr and Fr in the rice conditions), all the participants were asked to rate the accuracy of their food recognition (how much it looked like the target food) on a VAS before they started eating. Figure 3.9 shows the results. The averages (stdevs) for Rn and Fn were 39.4 (25.1) and 93.5 (7.5), respectively. Finally, the result for **Q7** on the ease of eating was ($M = 41.1, SD = 26.5$).

Discussion. From Fig. 3.7(a), it is marginally suggested that participants felt the taste of somen noodles more weakly in Fn than in Sn. Although not statistically significant, several participants also commented that they felt the taste of somen noodles more weakly in Rn than in Sn. These trends back up the findings in the first user study about the reduction of the perceived taste of the actual food that is eaten as a result of visual modulation. From Fig. 3.7(d), it was confirmed that they felt much more weakly that they were eating somen noodles in *R*n and *F*n than in *S*n. This result suggests that food recognition is more clearly manipulated by visual modulation than taste perception is.

The results of *S*n in Figs. 3.7(a) and (d) show that not all the participants felt that they were eating somen noodles or felt their taste even when they were eating and visually presented with somen noodles. Participants' comments in a post hoc interview include "It is typically served with cold soup so it was different from what I normally eat." and "I felt it was more like instant ramen noodles." These results suggest that some participants may have perceived the served somen noodles as something between typical somen and ramen noodles.

From Fig. 3.7(b), it was confirmed that participants felt the taste of ramen noodles more strongly in *R*n than in *F*n, and it is marginally suggested that they felt the taste of ramen noodles more strongly in *R*n than in Sn. These trends indicate that our StarGAN-based system can manipulate multisensory flavor perception and increase the perceived taste of food through its visual modulation. From Fig. 3.7(e), it was confirmed that participants felt more strongly that they were eating ramen noodles in *R*n than in *F*n and it is marginally suggested that they felt more strongly that they were eating ramen noodles in *R*n than in *S*n. These trends are also evidence that our system can manipulate food recognition.

In the first user study, all participants recognized the 3D model as ramen noodles. In the experiment, however, half of the participants did not think that StarGAN's output was ramen noodles in the case of *R*n. They first thought it was something else such as 'mentaiko spaghetti' (spicy cod roe pasta), 'roast beef rice bowl', or 'kanikama' (crab stick), which is presumably due to the pinkish color conversion. Note that the average scores and standard deviations of the results for **Q2** and **Q5** in the case of *S*n are relatively large, as shown in Figs. 3.7(b) and (e). This is probably also due to the fact that some participants recognized the served somen noodles as something between somen and ramen noodles.

From Fig. 3.7(c), it was confirmed that participants felt the taste of fried noodles much more strongly in *F*n than in *S*n and *R*n. From Fig. 3.7(f), it was confirmed that they felt much more strongly that they were eating fried noodles

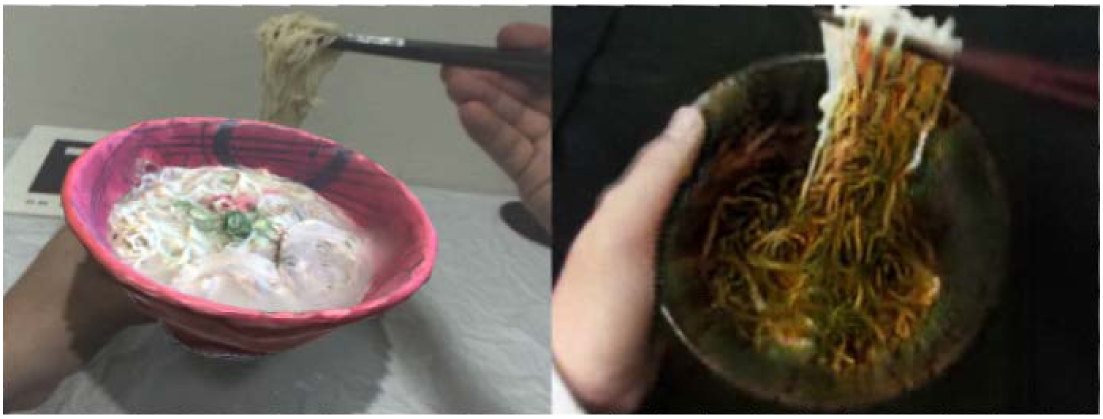


Figure 3.8: Left) The 3D model is static and it occludes the original food (somen noodles). Right) StarGAN dynamically generates a food image (fried noodles) that is visually coherent with the original food without occlusion.

in Fn than in Sn and Rn. These results again indicate that our StarGAN-based system can manipulate taste perception as well as food recognition through its visual modulation.

These results for the believability of fried noodles are more significant than those for ramen noodles. We believe this is related to the fact that all the participants thought that they were seeing fried noodles from the beginning in the case of Fn. That is, the strength of vision-induced gustatory manipulation greatly depends on the image quality.

Fried noodles are typically not served with soup, but the somen noodles in this experiment were. Interestingly, some participants gave comments such as “I perceived less soup” and “I did not notice any soup” in Fn. By exploiting the user’s preconceptions and high-quality visual modulation, it may even be possible to transfer foods between those with and without soup.

One participant mentioned that the food looked like fried noodles until he touched the food with the chopsticks and felt the sloshing of the soup. This episode reminds us of the importance of a multimodal interaction including tactile sensations.

3D model vs StarGAN. Comparing visual modulation by a 3D model and StarGAN, we found similar tendencies of the taste of the original food becoming

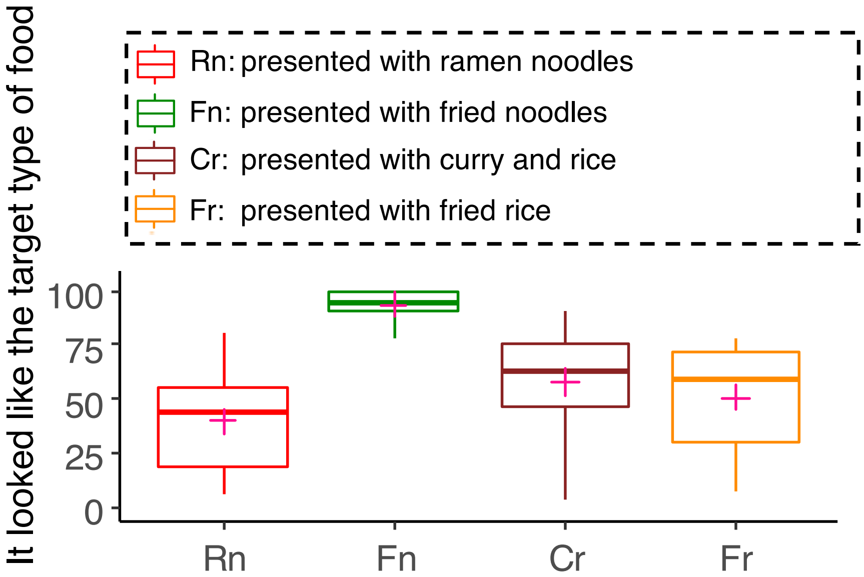


Figure 3.9: Results of the accuracy of participants' recognition of food. Before eating the food, participants answered on a VAS (0 and 100 being 'strongly disagree' and 'strongly agree', respectively).

weaker and that of the target food becoming stronger. However, the reduction or the gain is smaller in the case of StarGAN, which is presumably due to the lower image quality. In addition, it may also be due to the slightly larger latency.

As a positive aspect, the average ease of eating was increased from 27.6 to 41.1. Figure 3.8 illustrates typical views when a participant lifts some noodles with the chopsticks. In the case of StarGAN, the target food is visually coherent with the original food. Although a *t*-test did not find an statistically significant differences between StarGAN and the 3D model ($p = 0.12$), several participants of both experiments mentioned that it was clearly easier to eat when the StarGAN system was used, and no one gave the opposite opinion.

3.5.3 Rice Conditions

Results.

Hereinafter, we denote the steamed rice (original), curry and rice, and fried

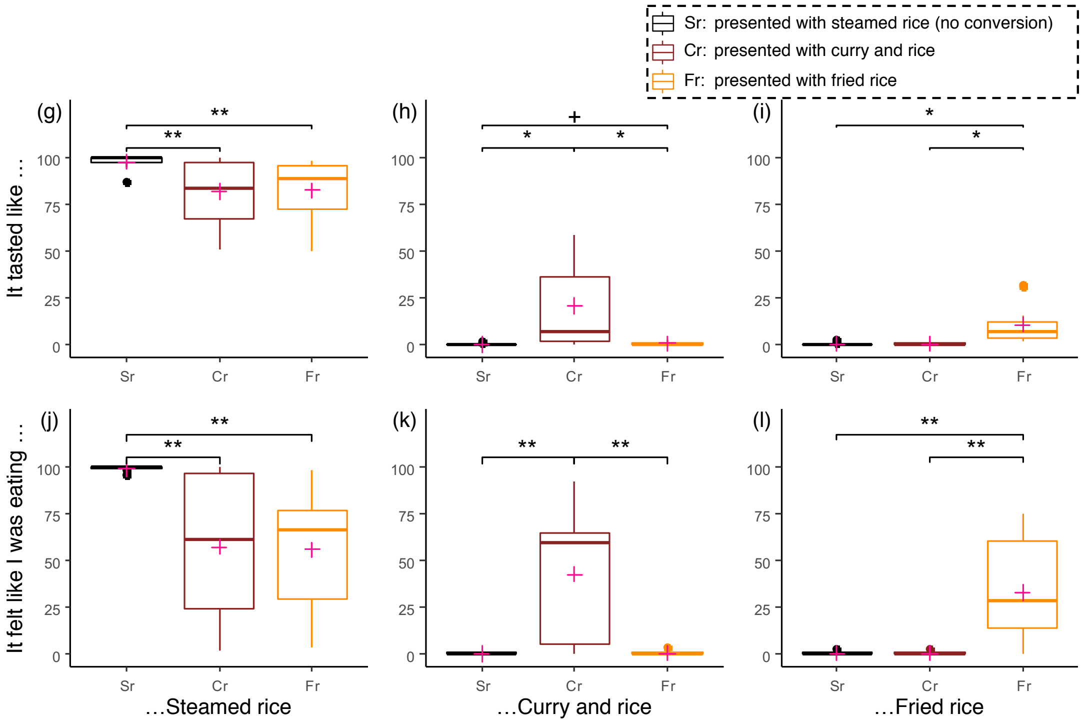


Figure 3.10: Results for rice conditions. The box plots in the upper and lower rows correspond to the results for **Q1** to **Q3** and **Q4** to **Q6**.

rice conditions by the symbols Sr, Cr, and Fr, respectively. Figure 3.10 shows the results of these rice conditions. The box plots in the upper and lower rows correspond to the results for **Q1** to **Q3** and **Q4** to **Q6**, respectively.

Fig. 3.10(g) shows the strength of the perceived taste of $s$teamed rice in the three visual modulation conditions Sr, Cr, and Fr. An ANOVA found an statistically significant differences ($F(2, 33) = 5.4274, p < 0.01$) and the Holm–Bonferroni method found statistically significant differences between Sr and Cr ($p < 0.01$) and between Sr and Fr ($p < 0.01$). Fig. 3.10(j) shows the confidence of the recognized food of $s$teamed rice in Sr, Cr, and Fr. An ANOVA found an statistically significant differences ($F(2, 33) = 8.4135, p < 0.01$) and the Holm–Bonferroni method found statistically significant differences between Sr and Cr ($p < 0.01$) and between Sr and Fr ($p < 0.01$).

Fig. 3.10(h) shows the strength of the perceived taste of $c$urry and rice in Sr, Cr, and Fr. An ANOVA found an statistically significant differences ($F(2, 33) =$

$8.8771, p < 0.001$) and the Holm–Bonferroni method found statistically significant differences between Sr and Cr ($p < 0.05$) and between Cr and Fr ($p < 0.05$), as well as a marginal trend toward significance between Sr and Fr ($p < 0.10$). Fig. 3.10(k) shows the confidence of the recognized food as $c$urry and rice in Sr, Cr, and Fr. An ANOVA found an statistically significant differences ($F(2, 33) = 17.631, p < 0.001$) and the Holm–Bonferroni method found statistically significant differences between Sr and Cr ($p < 0.01$) and between Cr and Fr ($p < 0.01$).

Fig. 3.10(i) shows the strength of the perceived taste of $f$ried rice in Sr, Cr, and Fr. An ANOVA found an statistically significant differences ($F(2, 33) = 11.731, p < 0.001$) and the Holm–Bonferroni method found statistically significant differences between Sr and Fr ($p < 0.05$) and between Cr and Fr ($p < 0.05$). Fig. 3.10(l) shows the confidence of the recognized food as $f$ried rice in Sr, Cr, and Fr. An ANOVA found an statistically significant differences ($F(2, 33) = 18.103, p < 0.001$) and the Holm–Bonferroni method found statistically significant differences between Sr and Fr ($p < 0.01$) and between Cr and Fr ($p < 0.01$).

In the case of Cr, 11 out of 12 participants (91.7%) thought that they were seeing curry and rice at the beginning. In the case of Fr, nine out of 12 participants (75%) thought that they were seeing fried rice at the beginning. The averages (stdevs) of the accuracy of participants' food recognition for Cr and Fr were 57.8 (25.8) and 50.7 (24.2), respectively (see Fig. 3.9). Finally, the result for **Q7** on the ease of eating was ($M = 29.7, SD = 17.8$).

Discussion. From Fig. 3.10(g), it was confirmed that participants felt the taste of steamed rice much more weakly in Cr and Fr than in Sr. From Fig. 3.10(j), it was confirmed that they felt much more weakly that they were eating steamed rice in Cr and Fr than in Sr. These results also demonstrate that our StarGAN-based visual modulation can weaken the perceived taste of the food that is actually eaten.

When presented with steamed rice without conversion (Sr), participants clearly thought that they were seeing and eating steamed rice. This is probably because there is little variation in its appearance and every participant is familiar with it. On the other hand, the large standard deviations for Cr and Fr in Fig. 3.10(j) indicate that for some participants, the perception of the food did not change very much.

From Fig. 3.10(h), it was confirmed that participamts felt the taste of curry and rice much more strongly in Cr than in St and Fr. From Fig. 3.10(k), it was confirmed that they felt much more strongly that they were eating curry and rice in Cr than in Sr and Fr. These results also show that our StarGAN-based visual modulation can strengthen the perceived taste of food that is visually presented.

Many participants commented that they thought they were eating *o*nly the rice part of the curry and rice in Cr. Interestingly, some participants thought it was different from normal steamed rice because they had the illusion of the flavor of curry spices. These results suggest that they tried to find a reason for the gap between the visual and gustatory stimuli either consciously or unconsciously from their past experience.

From Fig. 3.10(i), it was confirmed that participants felt the taste of fried rice much more strongly in Fr than in St and Cr. From Fig. 3.10(l), it was confirmed that they felt much more strongly that they were eating fried rice in Fr than in Sr and Cr. Again, these results indicate that our StarGAN-based visual modulation can strengthen the perceived taste of food that is visually presented.

However, the average scores of Fr in Figs. 3.10(i) and (l) are relatively low. This is probably because of the gap between the expected strong taste of fried rice and the actual near-tasteless steamed rice. Unlike the case of curry and rice, they could not conclude that they were eating only the rice part of fried rice. Sometimes the region around the bowl border was not modulated in our StarGAN-based image translation, which may also have contributed to the low scores.

The ease of eating was, on average, lower than that for the noodle conditions. This is probably because it is not common to eat steamed rice with a spoon. In addition, the cheap plastic spoon we used may have been too soft to easily eat the highly glutinous rice and it may have been difficult to tell the correct side of the spoon with the VST HMD.

3.5.4 Overall Discussion

Through the experiment, we confirmed that our StarGAN-based system successfully manipulated participants' multisensory flavor perception among multiple types of food.

The visual modulation was more effective in changing the food recognition of the participants. Many participants felt that they were eating the food that was visually presented rather than what they were actually eating. Note that they responded to **Q1 to Q6** not at the first glance before eating but after actually eating the food.

Interestingly, many participants also felt some changes in olfactory sensations ("I felt the smell and flavor of the food in my mouth."). We did not explicitly ask about the strength of the olfactory change, but our system seems to have cross-modal effects with other sensations in addition to the gustation.

For the noodle conditions, most participants answered in a post hoc interview that they felt the modulated taste for the second bite was as strong as that for the first bite. However, for the rice conditions, many participants saw through the illusion while eating. Typical comments indicating this include "I realized it was steamed rice after a while because of the lack of the unique spicy smell of curry and rice." and "It looked and tasted like curry and rice until I chewed it well." Some comments were more negative such as "I felt sad because it did not taste as I expected." and "I felt cognitive dissonance because what I expected and what I tasted did not match." This is probably due to the large gap between the near-tasteless steamed rice and the strong tastes of curry and rice and fried rice. Vision-induced gustatory manipulation will become less effective as the difference between the original and target tastes increases. These episodes also suggest that multisensory flavor perceptions dynamically change even when the visual modulation is the same, depending on the believability of other types of sensations, such as olfactory and haptics sensations.

Ernst and Banks reported that the human brain integrates visual and haptic information by considering the reliability of each type of information, but there seems to be a similar phenomenon in the multimodal interaction of multisensory flavor perceptions [36].

When the visual and gustatory stimuli conflicted, participants tried to find the reason. Other interesting comments in addition to those in Sec. 3.5.3 include "It tasted like some awful tasteless fried noodles thinned with hot water I once had." and "It tasted like some tasteless fried rice without seasoning that I made at home." Their multisensory flavor perceptions were influenced by their past experience and further investigation is necessary.

In relation to the memory recall of food experience, different numbers of participants recognized the modulated food correctly or incorrectly at first sight, depending on the food type. However, for all types of food we tested, no statistically significant differences was found between the correct and incorrect participants, probably due to the small number of participants. We may obtain some interesting results with more participants. In this experiment, we told the participants our intended target types of food; however, it will also be interesting to see the effects of their own impressions without telling them the answers.

Many participants gave positive comments on their general impression about vision-induced gustatory manipulation, such as "It is interesting and promising." "I see a future for it." and "It could be used to eat something that doesn't really exist like dragon meat." However, they also mentioned that the current system is not practical for daily use because of the cumbersome HMD and the resulting difficulty of eating.

3.6 Limitations

Even though we believe that our research is valuable overall, it has several limitations. In this section, we discuss some of these limitations and the future prospects for this technology.

3.6.1 Food Types and Participants

We used a limited number of original and target food types. The validity of the food selection is also arguable. Noodles and rice are very common in East Asian countries, and somen noodles and steamed rice were chosen as the original food types because they do not have strong tastes and are widely available. Target food types such as ramen noodles and fried rice were chosen because they are popular and often contain 'forbidden' ingredients. Thus, we claim that the selection is reasonable for our target application of supporting people who cannot eat what they want. However, our research did not indicate whether and how much we can manipulate multisensory flavor perceptions for other original and target food types. It will be interesting to investigate the relationship between the magnitude of manipulation and the similarity in appearance and/or taste between original and target types of food. The low geographical, ethnic, and gender variations among the participants are also possible limitations of this study. The results of the experiment, for example, would have been very different if the participants had been European. We would like to conduct follow-up studies using other combinations of food types with a larger variation of participants in the future.

3.6.2 Experimental Protocol

In our user studies, participants saw the original food at the beginning, and they knew what they were actually eating regardless of the visual modulation conditions. We intentionally chose this protocol because it is a natural assumption that a person is aware of the original food in actual food translation applications. However, in the literature on the perception of visual modulation, it is common not to reveal the original appearance to the participants. Our protocol seems to have reduced the strength of gustatory manipulation. Some participants commented that they would have felt the taste of the target food more strongly if they had not seen the original food in advance. In other words, it is expected that our system will be even more effective if the original food is not revealed, which we would like to confirm in the future.

In addition, the impact of the questionnaire on the results of the experiment also needs to be considered. The fixed order of the questions may have biased the experimental results. Being able to deduce the food that will be displayed next from the questionnaire may have also biased the results.

3.6.3 Image Translation Quality

There is considerable room for improvement in the quality of image translation. Food images collected from the Twitter stream are generally more colorful, taken near the dishes, have chopsticks or a spoon at similar angles, etc. These characteristics are different from those in the experiment, where the input images are taken from a slightly further distance and are relatively plain and colorless compared with Twitter stream images. This inconsistency may have sometimes caused translation failure. In addition, the current implementation does not extract a food region but translates the entire input image into the target food. This is a severe problem and we had to use white walls, a black desk, etc., to suppress unwanted results. Measuring the frequency of translation failure itself is also future work. We plan to develop a new neural network that can take food regions into consideration. Even when StarGAN worked well, its output was sometimes very different from the participants' expectation. This is because of a large variation within a single type of food (e.g., ramen noodles). Part of this problem may be addressed by also using CycleGAN or a similars GAN to tweak the results, e.g., the addition of toppings. We are confident that the image quality can be improved in the near future. In fact, we have already acquired better results with a new GAN architecture. Our new dataset now supports more than 100 types of food-to-food image translation, which will produce more natural and stable images under deformation. The image quality can also be improved by using pix2pixHD [124], vid2vid [125], etc., which are capable of video frame prediction.

3.6.4 System Latency

The end-to-end latency from image capture to the display of the translated image was around 300 to 500 [ms] with an average of around 400 [ms]. This is a very large latency for a modern AR system. Each of the three major steps, image transmission to the server, image-to-image translation, and transmission of the translated image to the client, takes about 100 [ms]. The total latency can be minimized without the necessity of network communication if both the client and server modules can run on a single machine. We have encountered resource and synchronization issues that prevented such an implementation. Note that despite the large latency, no participant appeared to have or reported related issues such as simulator sickness during the experiments. However, the latency should be minimized for more practical use. The apparent latency of the visual overlay should also be reduced. After the experiments, we implemented AR Timewarping [126] and heuristically confirmed that the apparent temporal registration error was reduced by 70 to 80%.

3.6.5 Difficulty of Eating

It was difficult to eat while wearing the HMD. The main reason is that it was difficult to recognize the food position due to the parallax between the camera and eye positions, the limited field of view of the HMD, and the system latency. Participants commented that they perceived their mouth position to be more forward than it actually was and that they could not see the food when it was being put into their mouths. StarGAN helped to improve the ease of eating to some extent; however, it should be improved further in the future. In this direction, we have already designed a HMD with a wider opening around the mouth and a smaller viewpoint offset, which we plan to test in follow-up studies.

3.7 Conclusion

In this chapter, we have described in detail a VST AR system for gustatory manipulation. Our system can flexibly change the appearance of one type of food into another in real time by using StarGAN-based image-to-image translation. Our experimental results revealed that our system successfully manipulates gustatory manipulations to some extent, but the effectiveness seems to depend on the original and target types of food as well as the experience of the individual with the food. Our research has several limitations, which also suggest interesting future directions. Regarding food types and participants, we used a limited number of food types, and the diversity and number of participants were both low. In the future, we will conduct follow-up studies using more types of food with a larger population and a more diverse demographic background. Regarding the image translation quality, we will develop a better neural network that can extract food regions and support a larger number of food types. Regarding the system latency, it was not problematic but noticeable. Since the experiments, we have already implemented AR Timewarping [126] and reduced apparent temporal registration errors, and we will use the revised system in follow-up studies in the future. Despite these limitations, we believe our research has shown, for the first time, the great potential of GAN-based cross-modal effects as a simple yet powerful tool for multimodal AR systems, and we hope many researchers will be encouraged to follow these promising trends.

4 Persistence of cross-modal effectse between vision and gustation

4.1 Gustatory Manipulation by GAN-based Visual Modulation Persists for Several Bites of Food

We have reported a gustatory manipulation interface that changes the perception of taste and type of food through visual modulation using GAN in Chap. 3. By changing the appearance of somen noodles to ramen noodles and fried noodles and steamed rice to curry and rice and fried rice, we found that the participants felt they were eating the food presented visually. On the other hand, some participants reported decreased multisensory flavor perception changes after eating the food more than once. Suppose the flavor change of the vision-induced gustatory manipulation interface does not last but only occurs in the first bite. In that case, it is challenging to apply it to daily meals where a lasting multisensory flavor perception change is necessary. To the best of our knowledge, no research focuses on multisensory flavor perception change persistence using cross-modal effects by visual modulation. In this chapter, we focused on the persistence of the multisensory flavor perception change.

In Chap. 3, participants directly viewed the unmodulated original food without wearing the HMD before the experiment. We hypothesize that looking only at the altered food without looking at the original one would reduce the belief about what they are eating, and thus the multisensory flavor perception change effect would increase.

Besides, Wan *et al.* suggest that the formation of cross-modal associations to taste differs depending on cultural differences such as the nationality and gender of the participants [59]. For example, cultural differences in nationality affect the cross-modal correspondence between visual features and taste [59, 60], and sound and taste [61]. The present study utilizes a cross-modal correspondence between food appearance and multisensory flavor perception, but does the effect differ across cultures based on nationality and gender? In this chapter, in addition to the persistence of multisensory flavor perception change, we investigated the effects of prior exposure to a specific food, nationality, and gender on the strength of multisensory flavor perception change.

In Chap. 3, we succeeded in changing the participants' perception of the taste and type of food in the first or second bite they ate. On the other hand, it is not clear whether the multisensory flavor perception change they experience lasts until they have eaten all their food (i.e., until the end of the meal). Some participants felt that eating food multiple times reduced the effect of multisensory flavor perception change. We aim to improve their QoL by using gustatory manipulation systems to alter the taste of the alternative foods they are eating, giving them the sensation of eating what they want. If they use the vision-induced gustatory manipulation interface daily, the gustatory manipulation's effect must last until their meal ends. However, despite its importance, there are no studies to our knowledge investigating the persistence of multisensory flavor perception change using cross-modal effects by visual modulation. Therefore, we measured the multisensory flavor perception change from the first to the fifth bite and investigated vision-induced gustatory manipulation persistence. Note that in this chapter, the term 'bite' is used to describe the series of actions from bringing a mouthful of food to the mouth, tasting it, and swallowing it. The biting or chewing actions themselves are not our focus of the study.

Besides, we also investigated "Would not seeing the original food before seeing the altered food affect the intensity of the gustatory manipulation, and how would it affect it?" and "Whether and how the intensity of the multisensory flavor perception change differ depend on the participants' cultural backgrounds (nationality and gender)" which had been problematic in previous studies. These contributions reveal that gustatory manipulation using visual modulation tends to be persistent for many users, indicating the potential for applications of gustatory manipulation interfaces.

The major contributions of this chapter are as follows:

- Vision-induced gustatory manipulation is persistent in many participants for several times of a mouthful of food. Their persistent multisensory flavor perception changes are divided into three groups: those in which the intensity of the multisensory flavor perception change gradually increased, those in which it gradually decreased, and those in which it did not fluctuate, each with about the same number of participants.
- Vision-induced gustatory manipulation is clearly present even when the original type of food was never shown to the participants directly without a HMD.
- Vision-induced gustatory manipulation is affected differently depending on the participant's attributes (gender and nationality). Still, those who are less familiar with the original and target types of food may have a stronger effect.

4.2 Experiment

4.2.1 Overview

The purpose of the experiment is to investigate the effectiveness of the GAN-based gustatory manipulation system from the following perspectives.

- Whether and how the gustatory manipulation persists while eating the modulated food.
- Whether and how not seeing the original food before seeing the modulated food affects the strength of the perceived multisensory flavor perception of the modulated food.
- Whether and how the results depend on the participants' nationality and gender.

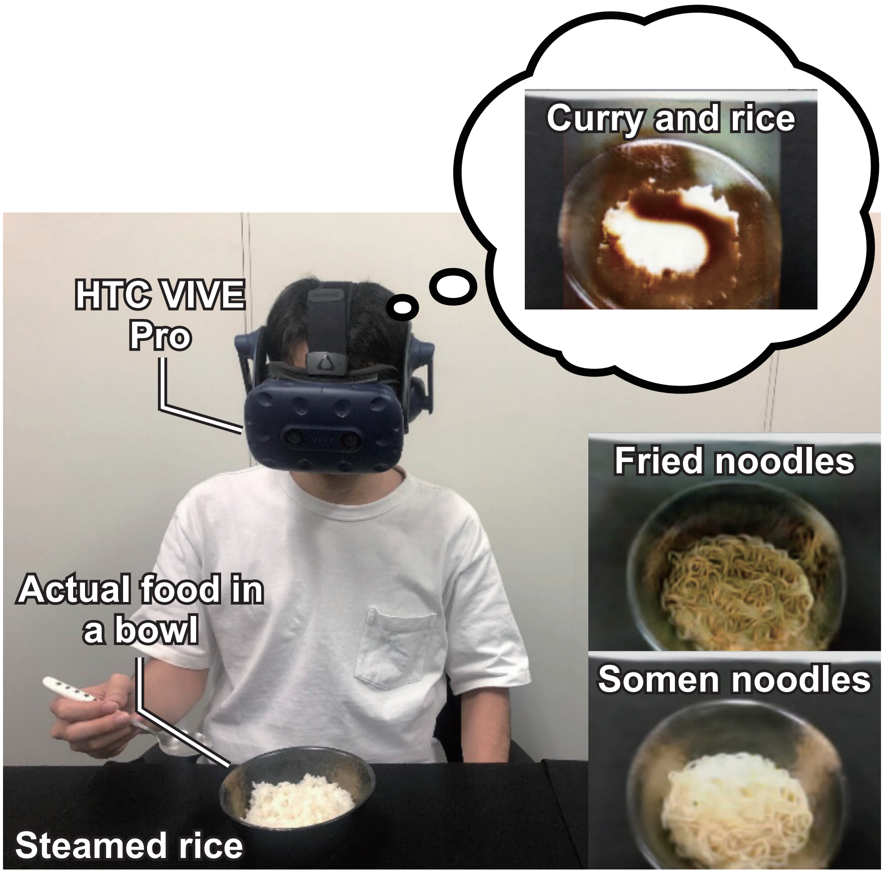


Figure 4.1: Our GAN-based real-time food-to-food translation system in action. Left) User with a VST HMD experiencing vision-induced gustatory manipulation of Rice Conditions. Right bottom) Examples of Noodle Conditions.

For the experiment, we reduced the actual and apparent motion-to-photon latency in the gustatory manipulation system. The actual latency was reduced from around 400 [ms] to around 150 [ms] by using a single PC to reduce the communication delay. The apparent latency (registration error) was further reduced by around 87% by implementing AR Timewarping [126].

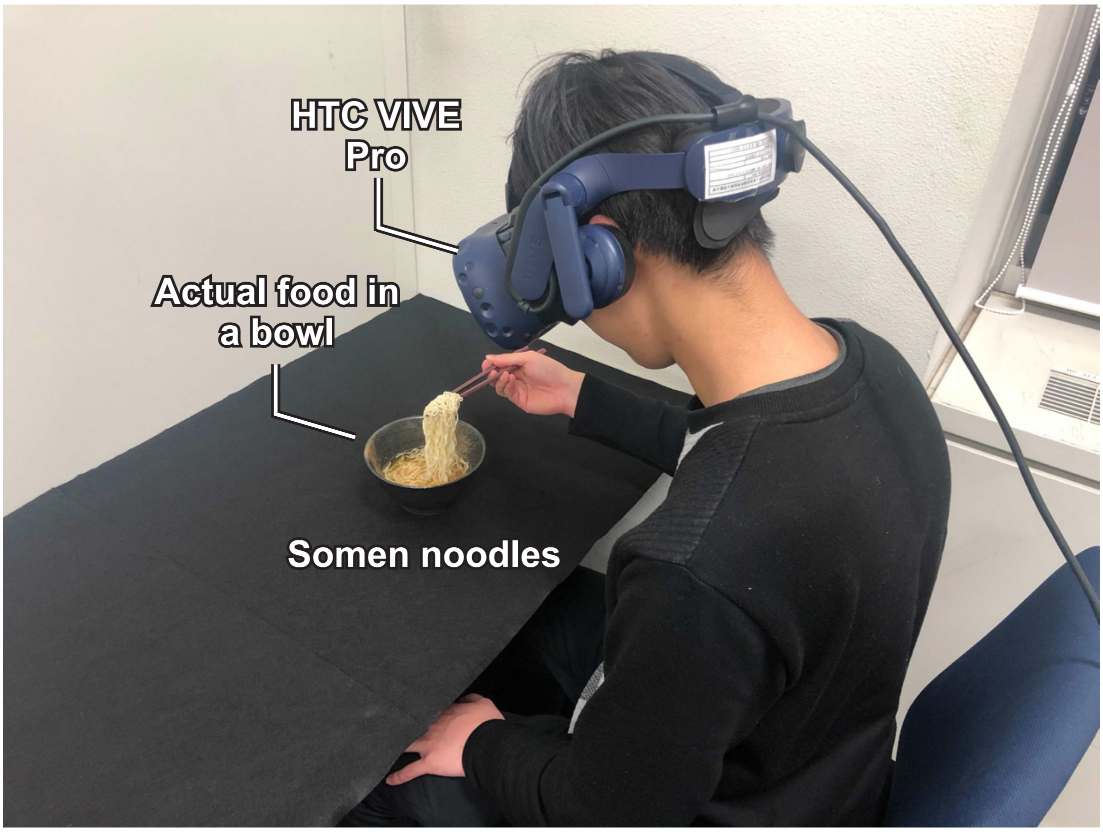


Figure 4.2: A participant in a noodle condition. They ate food placed on a black table surrounded by a white wall while wearing HMDs.

4.2.2 Procedure

Figures 4.1 and 4.2 show a participant in the experiment. The food was served in a black bowl with red chopsticks or a spoon on a black table in a quiet room with white walls for stable visual modulation. We confirmed that each participant was healthy and not too full or hungry. They wore the HMD and looked at either the original or the modulated food for three minutes to get used to the system and the viewing experience. Then, they were asked to tell what food they observed before eating. We told them the correct food type, and they answered whether or not it appeared so. Then, they took some water and a bite of the food, and answered the questions.

The questionnaire was displayed on the HMD so that participants could answer the questions through the HMD. They repeated the procedure five times with an interval of 120 [sec] per condition (about 60 [sec] for eating and answering, respectively) and finally removed the HMD.

As the number of trials increases, we reduced the number of food conditions from six [118] to four by removing the ramen noodles (**Rn**) and fried rice (**Fr**) conditions as they were less effective in Chap. 3 compared to the fried noodles (**Fn**) and curry and rice (**Cr**) conditions, respectively. Each participant performed all four conditions in a single day, either the noodle conditions (**Sn**, **Fn**) first or the rice conditions (**Sr**, **Cr**) first in a randomized order.

Following our previous study, the questionnaire consisted of the four questions below on a VAS (0 and 100 being 'strongly disagree' and 'strongly agree', respectively) [117]. **Q1** and **Q2** were about taste perception and **Q3** and **Q4** were about food recognition.

Q1. It tasted like *somen noodles* (or *steamed rice*).

Q2. It tasted like *fried noodles* (or *curry and rice*).

Q3. It felt like I was eating *somen noodles* (or *steamed rice*).

Q4. It felt like I was eating *fried noodles* (or *curry and rice*).

4.3 Results and Discussions

4.3.1 Overview

A total of 16 volunteers (eight males with an average age of 24.1 and eight females with an average age of 31.5 ranging from 22 to 49) participated in the experiment. None of them had previously tried our AR gustatory manipulation system. They were briefed about the procedure and its purpose orally and agreed to it. The institutional review board approved the experiment. The participants were eight Japanese and eight non-Japanese (two French, two Thai, one Chinese, one German, one Korean, and one Egyptian) recruited from our university. The participants had eaten all of the four types of food used in the experiment before the experiment. In the following, we give the results and discussion for the noodle and rice conditions in order. We performed a three-way ANOVA followed by a post hoc analysis with the Holm–Bonferroni correction throughout the experiment. As our data were not normally distributed, we employed the aligned rank transform procedure for hypothesis testing [127]. However, it should be noted that the number of participants participating In this chapter was by no means large, and the experimental results should be associated with the individual differences. Nevertheless, we believe the experimental results will give interesting insights.

4.3.2 Noodle Conditions

Results

The results of the noodle conditions are shown in Figure 4.3. The results for **Q1** and **Q2**, **Q3** and **Q4**, are shown in the box plots in the upper and lower rows, respectively. **SnTaF** in Figure 4.3 denotes, for example, the intensity of the perceived **taste** of *fried noodles* when somen noodles were presented visually (the visual modulation condition *Sn*). Additionally, **FnTyS** in Figure 4.3 denotes the intensity of the perceived **type** of *somen noodles* when fried noodles were presented visually (the visual modulation condition *Fn*). Sni denotes the result for the i-th bite showing the persistency trend.

Paired t-tests between **Sn1** and **Fn1** found significant differences for all groups ($p < 0.01$ for **SnTaS** and **FnTaS**, $p < 0.05$ for **SnTaF** and **FnTaF**, $p < 0.001$ for **SnTyS** and **FnTyS**, and $p < 0.01$ for **SnTyF** and **FnTyF**). For example, in the upper left graph of Figure 4.3, which shows the results for "It tasted like somen noodles", **Fn1** is lower than **Sn1**. This means that the visual modulation changed the food's appearance from the original somen noodles into fried noodles, and the taste of the original food, somen noodles, was more weakly perceived. In the upper right graph of Figure 4.3, which shows the result of "It tasted like fried noodles", **Fn1** is higher than **Sn1**. This means that the visual modulation changed the food's appearance into that of fried noodles and the taste of fried noodles was more strongly perceived. These visual modulation trends decreasing the taste of the original food and increasing the taste of the visually presented food are similar to the previous studies. Additionally, the lower graphs of Figure 4.3,

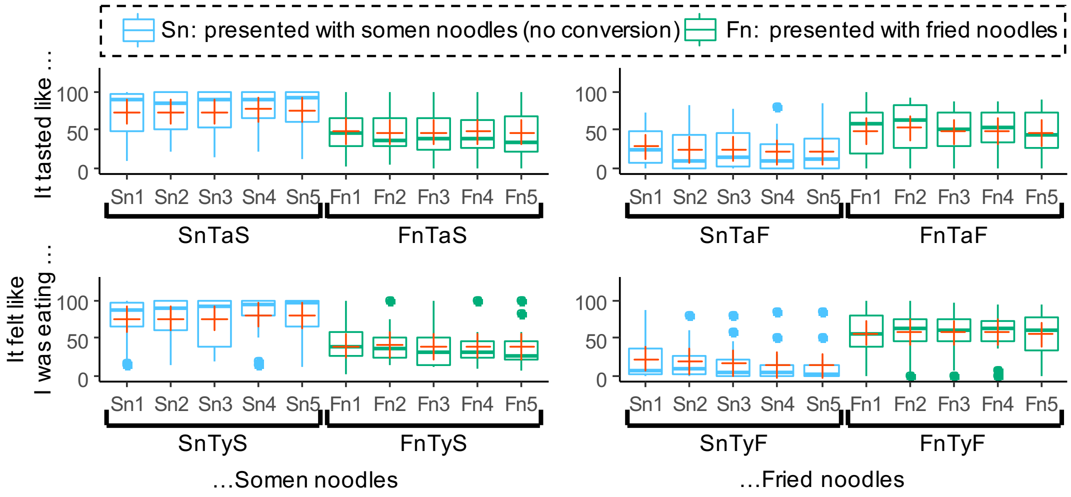


Figure 4.3: VAS scores in the noodle conditions. The box plots in the upper and lower rows correspond to the results for **Q1** and **Q2**, and **Q3** and **Q4**, respectively. The red crosses indicate the mean values, and the dots indicate the outliers. For example, the upper left graph, where "It tasted like . . . " and ". . . Somen noodles" intersect, shows the results of Q1, where "It tasted like somen noodles" was asked, and **Sn1** shows the VAS score of the first bite under **Sn** conditions. In addition, the entire **Sn1**–**Sn5** is denoted as **SnTaS** for identification purposes.

which correspond to the results of "It felt like I was eating somen noodles" and "It felt like I was eating fried noodles", show more significant differences between **Sn1** and **Fn1** compared to the upper graphs. The results are similar to those of the previous studies, showing that vision-induced gustatory manipulation is more effective in changing participant's perception of the food that they are eating than changing their perception of taste. Similarly to the previous user study, gustatory manipulation is clearly present in the noodle conditions.

Table 4.1 shows the number of participants whose VAS scores have changed by 10 points or more between the first and fifth bites under the noodle conditions. We call the groups with increasing and decreasing scores **Up** and **Down**, respectively, and the group with little changing scores **Stay** . Figure 4.4 shows the relative change in VAS scores from the first bite to the fifth bite in each

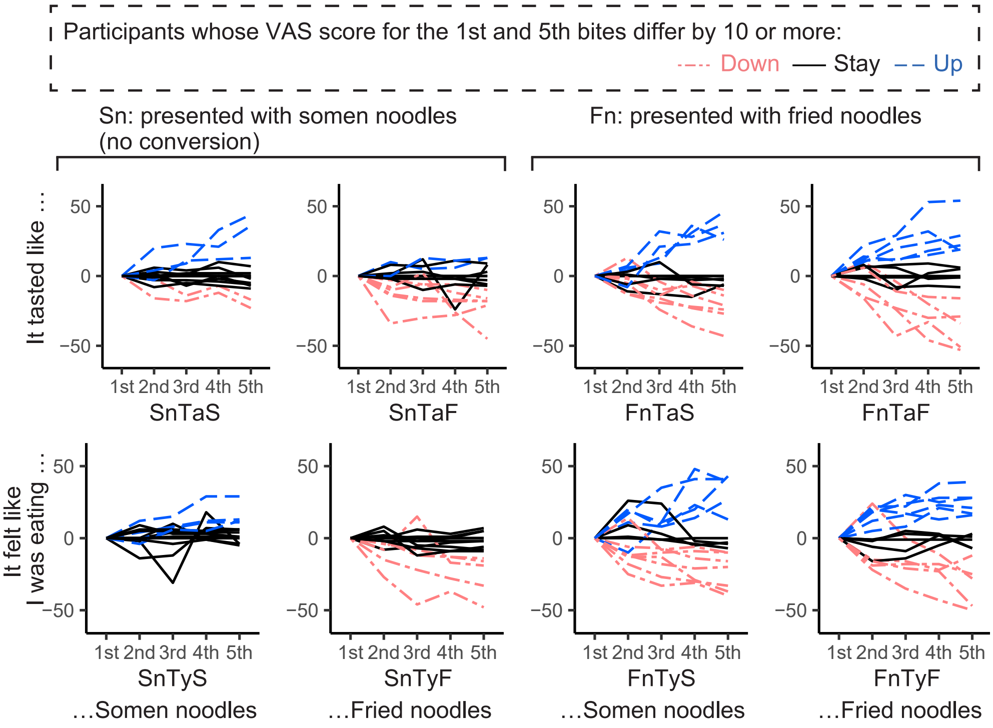


Figure 4.4: Relative change in VAS scores from the first to the fifth bite for each participant. Participants were divided into three groups based on the differences in VAS scores between the first and fifth bites: **Up** group (those increased by 10 points or more), **Down** group (those decreased by 10 points or more), and **Stay** group (those changed by less than 10 points). For example, the upper left graph, where "It tasted like. . . " and ". . . Somen noodles" intersect, shows the results of **SnTaS**, where "It tasted like somen noodles" was asked, and the n-th VAS score shows the VAS score of the n-th bite under Sn conditions.

participant's noodle conditions, with the first bite as the baseline. The results for each participant were categorized as **Up**, **Down**, and **Stay** according to Table 4.1. Our hypothesis was that "the cross-modal effect would decrease as the number of bites increased, and participants would feel more strongly that they were eating the original food." For example, the values of **SnTaS** and **FnTaS** were expected to increase as the number of bites increases, while the values of **SnTaF** and **FnTaF** were expected to decrease.

Tables 4.2 and 4.3 show the results of ANOVA for **Q1** to **Q4** for each type of the noodles. Here, the interactions of “Persistency: Nationality”, “Persistency: Gender” and “Persistency: Nationality: Gender” are omitted in the tables because there was no significant difference. Significant differences are indicated with symbols (*** for $p < 0.001$, ** for $p < 0.01$, * for $p < 0.05$, and + for $p < 0.1$). Figures 4.5 and 4.6 show the results of nationality and gender differences. If the score for “It tasted like somen noodles” was lower for **Fn** than **Sn**, and the score for “It tasted like fried noodles” was higher for **Fn** than **Sn**, the effect of visual modulation was stronger. For example, if the values of **SnTaS** and **FnTaF** were large and the values of **SnTaF** and **FnTaS** were small in Figure 4.5, the effect of visual modulation was considered to be strong.

Discussion

Figure 4.3 shows the same tendency as the results of the noodle conditions in Chap. 3: visual modulation decreases the taste of the original food and increases the taste of the visually presented food. However, in Figure 4.3, **SnTaF** and **SnTyF** showed higher scores than those in Chap. 3, which were close to zero,

Table 4.1: Number of participants whose VAS scores for the first and fifth bites differ by 10 or more in the noodle conditions (up: number of participants with improved VAS scores, stay: number of participants with little change in VAS scores, down: number of participants with reduced VAS scores).

	Sn				**Fn**			
	taste		**type**		**taste**		**type**	
	SnTaS	**SnTaF**	**SnTyS**	**SnTyF**	**FnTaS**	**FnTaF**	**FnTyS**	**FnTyF**
up	3	2	4	0	4	5	5	6
stay	11	8	12	11	6	6	4	5
down	2	6	0	5	6	5	7	5

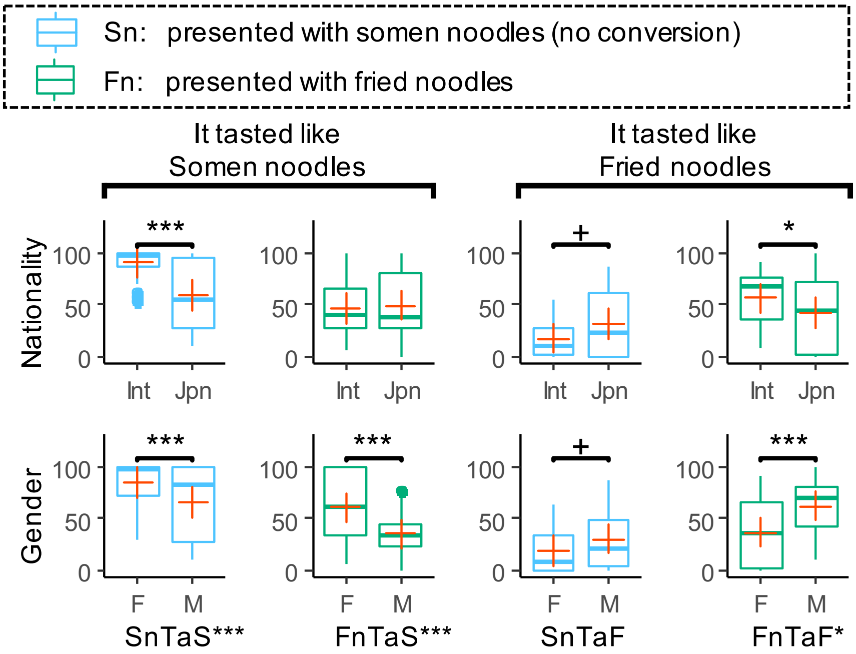


Figure 4.5: VAS scores with regard to nationality and gender for **Q1** and **Q2** in the noodle conditions. Interactions are indicated to the right of result group names **SnTaS**, **FnTaS**, and **FnTaF**. The red crosses indicate the mean values, and the dots indicate the outliers. The upper graphs show the nationality classification, where Int indicates international participants and Jpn indicates Japanese participants. The lower graphs show the classification of gender, where F indicates female participants and M indicates male participants. For example, the upper left graph shows the values of **Sn1**–**Sn5** of **SnTaS** focusing on nationality.

indicating that some participants felt that the food they were eating was fried noodles even under the somen noodle condition. A possible cause for this problem is that some participants experienced the **Fn** condition before the **Sn** condition (without knowing what the original food was), which have affected the multisensory flavor perception. However, when presented with somen noodles, the **Up** group is larger than the **Down** group in **SnTaS** and **SnTyS**, and the **Down** group is larger than the **Up** group in **SnTaF** and **SnTyF** in Table 4.1

and Figure 4.4. This appears to mean that the participants gradually became more confident that they were eating somen noodles when visually presented with

Table 4.2: ANOVA results for **Q1** and **Q2** in the noodle conditions. "Nationality: Gender" shows the interaction between nationality and gender. The interactions of "Persistency: Nationality", "Persistency: Gender," and "Persistency: Nationality: Gender" are omitted because there was no significant difference.

	It Tasted Like. . .			
	. . . Somen Noodles		. . . Fried Noodles	
Condition	**SnTaS**	**FnTaS**	**SnTaF**	**FnTaF**
Persistency	n.s.	n.s.	n.s.	n.s.
Nationality	***	n.s.	+	*
Gender	***	***	+	***
Nationality: Gender	***	***	n.s.	*

Table 4.3: ANOVA results for **Q3** and **Q4** in the noodle conditions. "Nationality: Gender" shows the interaction between nationality and gender. The interactions of "Persistency: Nationality", "Persistency: Gender," and "Persistency: Nationality: Gender" are omitted because there was no significant difference.

	It Felt Like I Was Eating. . .			
	. . . Somen Noodles		. . . Fried Noodles	
Condition	**SnTyS**	**FnTyS**	**SnTyF**	**FnTyF**
Persistency	n.s.	n.s.	n.s.	n.s.
Nationality	***	n.s.	*	*
Gender	***	n.s.	*	n.s.
Nationality:Gender	***	**	n.s.	n.s.

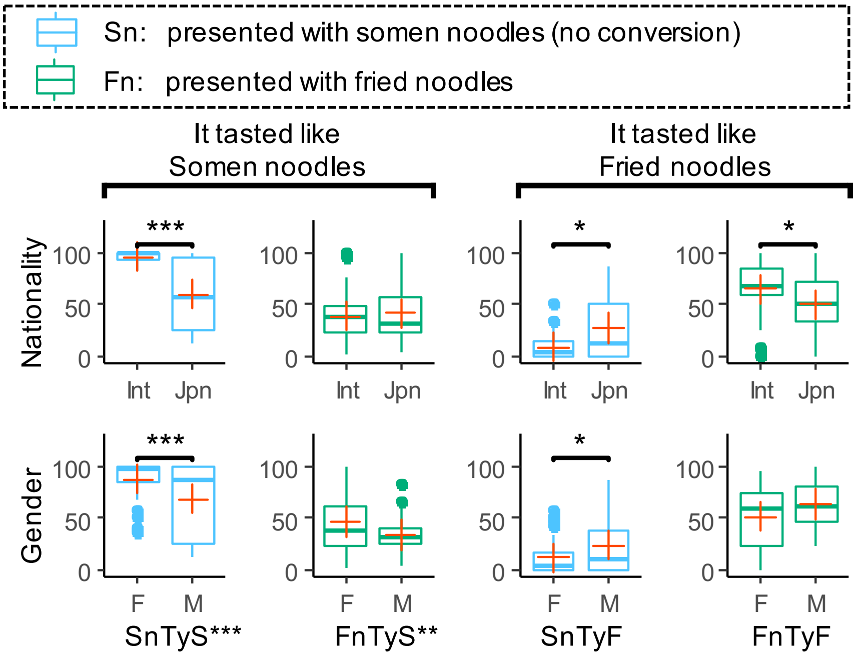


Figure 4.6: VAS scores with regard to nationality and gender for **Q3** and **Q4** in the noodle conditions. Interactions are indicated to the right of result group names **SnTyS** and **FnTyS**. The red crosses indicate the mean values, and the dots indicate the outliers. The upper graphs show the nationality classification, where Int indicates international participants and Jpn indicates Japanese participants. The lower graphs show the classification of gender, where F indicates female participants and M indicates male participants. For example, the upper left graph shows the values of **Sn1**–**Sn5** of **SnTyS** focusing on nationality.

somen noodles.

No significant difference was found in all conditions on persistency in Tables 4.2 and 4.3 suggesting that the gustatory manipulation persists to some extent for a longer period of time. On the other hand, as can be read from Table 4.1, we could confirm that some participants' taste perceptions changed between the first and fifth bites. Focusing on **FnTaF** and **FnTyF**, which were the plausibility of the perceived taste and the recognized food type as fried noodles when visually presented with fried noodles, nearly the same number of the participants are classified into **Up** and **Down** groups, leaving the similar number of participants whose scores did not change much.

Additionally, Figure 4.4 shows that there were more participants in **Up** groups in **SnTaS** and **SnTyS** and more in **Down** groups in **SnTaF** and **SnTyF** with larger changes. These results indicated that even if the participants initially misidentified that they were eating fried noodles, they tended to recognize that they were eating somen noodles as the number of bites increased. In other words, it supported the hypothesis that "the cross-modal effect would decrease as the number of bites increased, and people would feel more strongly that they were eating the original food." On the other hand, **FnTaS**, **FnTyS**, **FnTaF**, and **FnTyF** in the **Fn** conditions had similar participant numbers in the Up, Stay, and **Down** groups, confirming that they were near evenly distributed. In particular, it is important to note that the **Down** groups in **FnTaS**, **FnTyS**, **FnTaF** and **FnTyF** differed from the trend in the **Sn** conditions in that there were also similar number of participants in the opposite (**Up**) groups. In other words, a non-negligible number of participants felt that they were eating fried noodles with a stronger confidence as the number of bites increased. These results differ from the hypothesis and indicate that the cross-modal effect of visual modulation was increased by multiple bites for some individuals.

We considered that the temporal change of the vision-induced gustatory manipulation effect varied between individuals. From these results, we could confirm that vision-induced gustatory manipulation was persistent in many participants. Their persistent multisensory flavor perception changes were divided into three groups: those in which the intensity of the multisensory flavor perception change gradually increased, those in which it gradually decreased, and those in which it did not fluctuate, each with about the same number of participants.

Next, we investigate evaluation scores in terms of nationality and gender. From Tables 4.2 and 4.3, we could confirm significant differences and trends toward significance in many groups. Looking at the nationality rows in Figures 4.5 and 4.6, the international participants felt the stronger taste of somen noodles when visually presented with somen noodles than the Japanese participants (**SnTaS** and **SnTyS**). We believe that this is because the foreign participants did not have much experience with somen noodles, and they may not have been confident in the taste of somen noodles. Note that the somen noodles used in this experiment were served not with typical cold soup but with warm soup. We adopted a ready-made instant somen noodles ** product with warm soup to make it consistent with hot fried noodles in this experiment. The Japanese participants have felt like they were eating fried noodles more strongly than the international participants under the **Sn** condition (**SnTaF** and **SnTyF**). We believe that this is because they did not feel like they were eating (typical cold) somen noodles.

Besides, the international participants felt the stronger taste of fried noodles when visually presented with fried noodles than the Japanese participants (**FnTaF** and **FnTyF**). Again, we believe that this is because the international participants had little experience of eating fried noodles and had a narrower range of expectations about the taste of fried noodles. Looking at the gender rows of Figures 4.5 and 4.6, the female participants felt the stronger taste of somen noodles when visually presented with somen noodles (**SnTaS** and **SnTyS**) and the weaker taste of fried noodles when visually presented with fried noodles (**FnTaF**) than the male participants. We believe that this result suggests that cultural differences such as cooking experience and average age (24.1 for males vs. 30.5 for females) affected the formation of cross-modal associations with taste [59]. However, in the **Fn** condition (**FnTyF**), there was no significant difference in the change in food type between the female and male groups. This result suggests little difference between men and women in the perception of food types after visual modulation. Despite these statistically significant differences, we have to also note that the generalizability of our findings is limited due to the small number of participants.

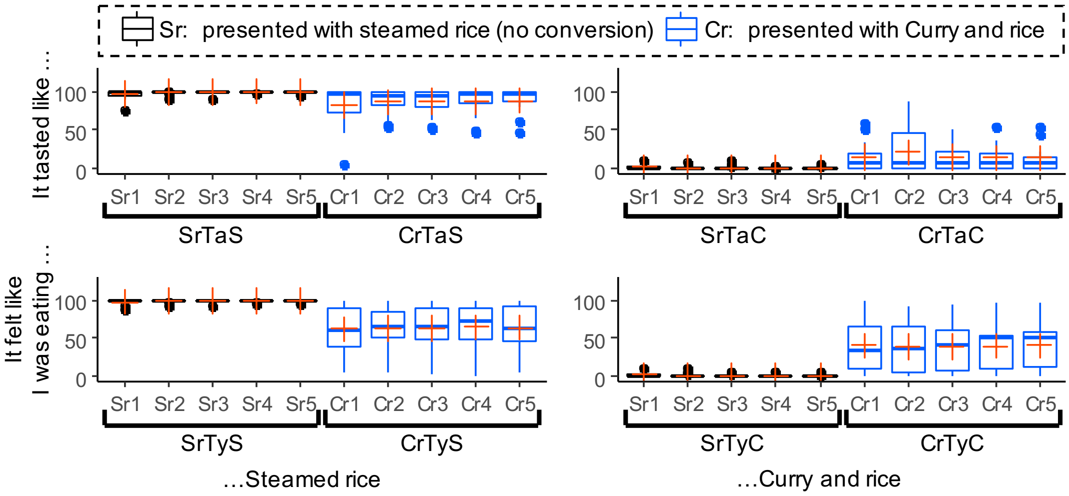


Figure 4.7: VAS scores in the rice conditions. The box plots in the upper and lower rows correspond to the results for **Q1** and **Q2**, and **Q3** and **Q4**, respectively. The red crosses indicate the mean values, and the dots indicate the outliers. For example, the upper left graph, where "It tasted like ..." and "...Steamed rice" intersect, shows the results of Q1, where "It tasted like steamed rice" was asked, and **Sr1** shows the VAS score of the first bite under **Sr** conditions. In addition, the entire **Sr1**–**Sr5** is denoted as **SrTaS** for identification purposes.

4.3.3 Rice Conditions

Results

Figure 4.7 shows the results of the rice conditions. The box plots in the upper and lower rows correspond to the results for **Q1** and **Q2**, and **Q3** and **Q4**, respectively. For example, in Figure 4.7, **SrTaC** denotes the strength of the perceived **taste** of *curry and rice* when visually presented with steamed rice (the visual modulation condition **Sr**). Additionally, in Figure 4.7, **CrTyS** denotes the strength of the perceived **type** of *steamed rice* when visually presented with curry and rice (the visual modulation condition **Cr**). Paired t-tests between **Sr1**

[**] Quality Labeling Standards for Dried Noodles, `https://www.cao.go.jp/consumer/history/01/kabusoshiki/syokuhinhyouji/doc/004_101004_shiryou2-5.pdf`, last accessed March 17, 2023.

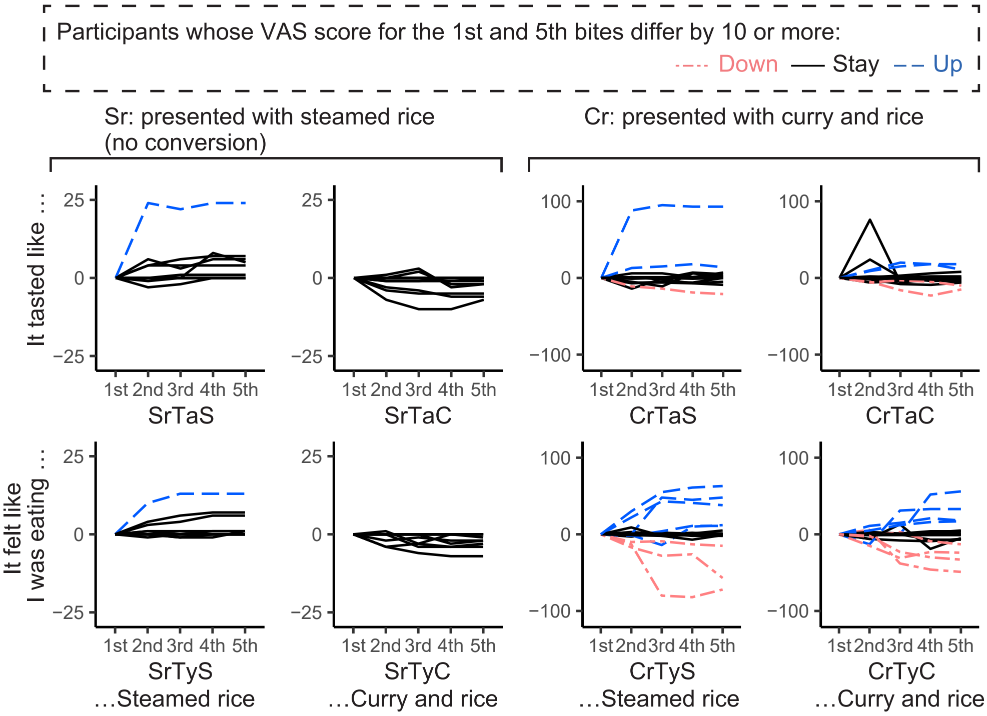


Figure 4.8: Relative change in VAS scores from the first to the fifth bite for each participant. Participants were divided into three groups based on the difference in VAS scores between the first and fifth bites: **Up** group (those increased by 10 points or more), **Down** group (those decreased by 10 points or more), and **Stay** group (those changed by less than 10 points). For example, the upper left graph, where "It tasted like . . . " and ". . . Steamed rice" intersect, shows the results of **SrTaS**, where "It tasted like steamed rice" was asked, and the n-th VAS score shows the VAS score of the n-th bite under **Sr** conditions.

and **Cr1** found significant differences for all groups ($p < 0.05$ for **SrTaS** and **CrTaS**, $p < 0.05$ for **SrTaC** and **CrTaC**, $p < 0.001$ for **SrTyS** and **CrTyS**, and $p < 0.001$ for **SrTyC** and **CrTyC**). For example, in the upper left graph of Figure 4.7, which shows the results for "It tasted like steamed rice", **Cr1** is lower than **Sr1**. This means that the visual modulation changed the food's appearance from the original steamed rice into curry and rice, and the taste of the original food, steamed rice, was more weakly perceived. In the upper right graph of Figure 4.7, which shows the result of “It tasted like curry and rice”, **Cr1** is higher than **Sr1**. This means that the visual modulation changed the food’s appearance into that of curry and rice and the taste of curry and rice was more strongly perceived. These visual modulation trends decreasing the taste of the original food and increasing the taste of the visually presented food are similar to the findings of the previous study and the noodle conditions above. Additionally, the lower graphs of Figure 4.7, which correspond to the results of “It felt like I was eating steamed rice” and “It felt like I was eating curry and rice”, show more significant differences between **Sr1** and **Cr1** compared to the upper graphs. The results are similar to those of the previous study and the noodle conditions above, showing that vision-induced gustatory manipulation is more effective in changing participant’s perception of the food that they are eating than changing their perception of taste. Similarly to the previous user study, gustatory manipulation is clearly present in the rice conditions as well.

Table 4.4 shows the number of participants whose VAS scores changed by 10 points or more between the first and fifth bites under the rice conditions. Figure 4.8 shows the relative change in VAS scores from the first bite to the fifth bite in each participant’s rice conditions, with the first bite as the baseline. The results for each participant were categorized as **Up**, **Down**, and **Stay** according to Table 4.4. Our hypothesis was that “the cross-modal effect would decrease as the number of bites increased, and participants would feel more strongly that they were eating the original food.” For example, the values of **SrTaS** and **CrTaS** were expected to increase as the number of bites increases, while the values of **SrTaC** and **CrTaC** were expected to decrease.

Tables 4.5 and 4.6 show the results of ANOVA for **Q1** to **Q4** for each of the rice conditions. The interactions of “Persistency: Nationality”, “Persistency: Gender,” and “Persistency: Nationality: Gender” are omitted because there was no significant difference. Figures 4.9 and 4.10 show the results of nationality and gender differences. If the score for “It tasted like steamed rice” is lower for **Cr** than **Sr**, and the score for “It tasted like curry and rice” is higher for **Cr** than **Sr**, the effect of visual modulation is stronger. For example, if the values of **SrTaS** and **CrTaC** are large and the values of **SrTaC** and **CrTaS** are small in Figure 4.7, the effect of visual modulation is considered to be strong.

Table 4.4: Number of participants whose VAS scores for the first and fifth bites differ by 10 or more in the rice conditions (up: number of participants with improved VAS scores, stay: number of participants with no change in VAS score, down: number of participants with reduced VAS scores).

	Sr				**Cr**			
	taste		**type**		**taste**		**type**	
	SrTaS	**SrTaC**	**SrTyS**	**SrTyC**	**CrTaS**	**CrTaC**	**CrTyS**	**CrTyC**
up	1	0	1	0	2	2	5	4
stay	15	16	15	16	13	12	8	8
down	0	0	0	0	1	2	3	4

Table 4.5: ANOVA results for **Q1** and **Q2** in the rice conditions. "Nationality: Gender" shows the interaction between nationality and gender. The interactions of "Persistency: Nationality", "Persistency: Gender," and "Persistency: Nationality: Gender" are omitted because there was no significant difference.

	It Tasted Like…			
	…Steamed Rice		…Curry and Rice	
Condition	**SrTaS**	**CrTaS**	**SrTaC**	**CrTaC**
Persistency	**	n.s.	n.s.	n.s.
Nationality	*	***	+	***
Gender	+	***	n.s.	***
Nationality:Gender	n.s.	***	n.s.	**

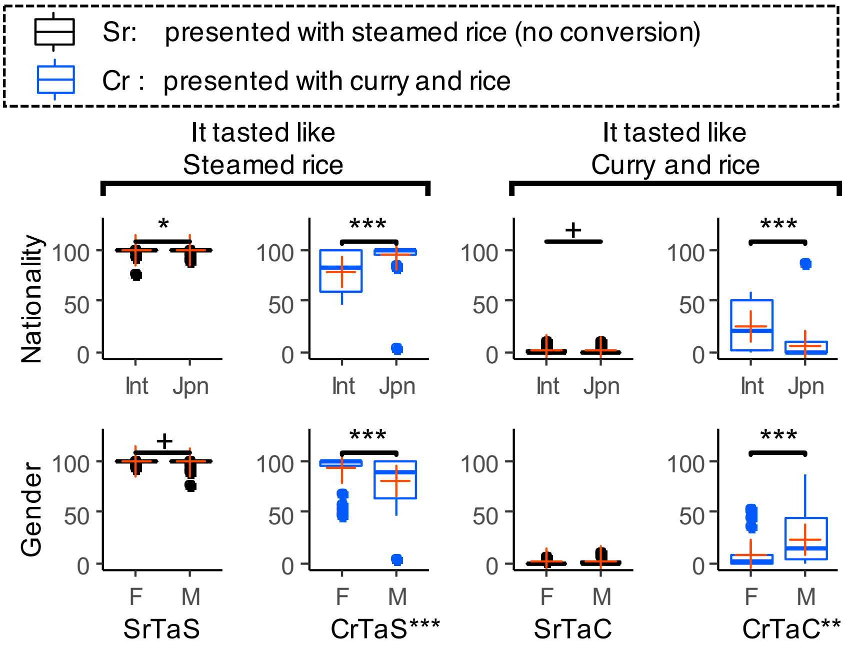


Figure 4.9: VAS scores with regard to nationality and gender for **Q1** and **Q2** in the rice conditions. Interactions are indicated to the right of result group names **CrTaS** and **CrTaC**. The red crosses indicate the mean values, and the dots indicate the outliers. The upper graphs show the nationality classification, where Int indicates international participants and Jpn indicates Japanese participants. The lower graphs show the classification of gender, where F indicates female participants and M indicates male participants. For example, the upper left graph shows the values of **Sn1**–**Sn5** of **SrTaS** focusing on nationality.

Discussion

Figure 4.7 shows similar scores and tendencies as the results of the rice conditions in Chap. 3: visual modulation decreases the taste of the original food and increases the taste of the visually presented food. As in the noodle conditions, multisensory flavor perceptions are manipulated successfully even if the participants only saw the evaluating food through the HMD.

Regarding persistency in Tables 4.5 and 4.6, a significant difference was con-

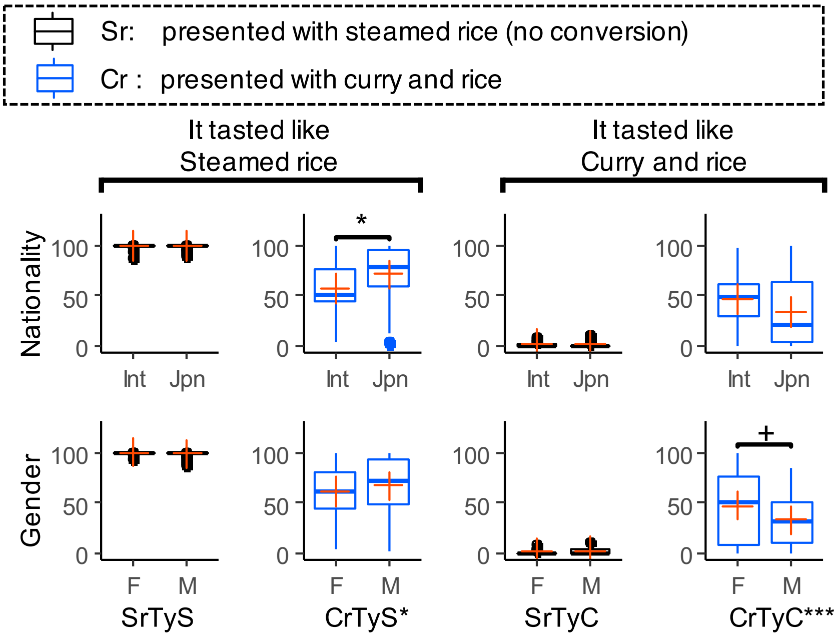


Figure 4.10: VAS scores with regard to nationality and gender for **Q3** and **Q4** in the rice conditions. Interactions are indicated to the right of result group names **CrTyS** and **CrTyC**. The red crosses indicate the mean values, and the dots indicate the outliers. The upper graphs show the nationality classification, where Int indicates international participants and Jpn indicates Japanese participants. The lower graphs show the classification of gender, where F indicates female participants and M indicates male participants. For example, the upper left graph shows the values of **Sn1**–**Sn5** of **SrTyS** focusing on nationality.

firmed only for **SrTaS**. However, no significant difference was found in the post-hoc analysis using the Holm method. For **SrTaS**, most participants scored 90 or higher, and only one participant changed the score by 10 or more, as shown in Table 4.4. From these results, we can say that the gustatory manipulation clearly persists in the rice conditions as well. Additionally, Figure 4.8 shows that most participants' scores did not change in the **Sr** conditions **SrTaS**, **SrTaC**, **SrTyS**, and **SrTyC**. Besides, **CrTaS** and **CrTaC**, which are the results of the questions on taste in the **Cr** conditions, showed little change in taste, which is different from the results for the noodle conditions. We hypothesize that these results are because steamed rice is an everyday food in Japan, and therefore the memory of the taste is robust, making it difficult to induce cross-modal effects and that the light taste of steamed rice makes it difficult to perceive changes in taste.

On the other hand, **CrTyS** and **CrTyC**, which are the results of the questions on food type in the **Cr** conditions, showed that the numbers of participants in the **Up** and **Down** groups were higher, confirming the similar trends as in the noodle conditions. These results indicate that as the number of bites increased, some participants noticed that they were eating steamed rice and the effect of the illusion decreased, while others thought they were eating curry and rice, and the effect of the illusion increased. We attribute the differences in taste and food type changes in these **Cr** conditions to the fact that the vision-induced gustatory manipulation is more effective in changing participants' perception of food type than changing their perception of taste. From these results, we can confirm again that the vision-induced gustatory manipulation is persistent in many participants. Their persistent multisensory flavor perception changes are divided into three

Table 4.6: ANOVA results for **Q3** and **Q4** in the rice conditions. "Nationality: Gender" shows the interaction between nationality and gender. The interactions of "Persistency: Nationality", "Persistency: Gender," and "Persistency: Nationality: Gender" are omitted because there was no significant difference.

	It Felt Like I Was Eating…			
	…Steamed Rice		…Curry and Rice	
Condition	**SrTyS**	**CrTyS**	**SrTyC**	**CrTyC**
Persistency	n.s.	n.s.	n.s.	n.s.
Nationality	n.s.	*	n.s.	n.s.
Gender	n.s.	n.s.	n.s.	+
Nationality:Gender	n.s.	*	n.s.	***

groups: those in which the intensity of the multisensory flavor perception change gradually increased, those in which it gradually decreased, and those in which it did not fluctuate, each with about the same number of participants.

Next, we investigate evaluation scores in terms of nationality and gender. From Tables 4.5 and 4.6, we can confirm significant differences and trends toward significance in many groups. In addition, the interaction between nationality and gender has also been confirmed. Looking at the nationality rows in Figures 4.9 and 4.10, the international participants felt the stronger taste of curry and rice when visually presented with curry and rice than Japanese participants like in the rice conditions (**CrTaC** and **CrTyC**). Again, we consider this is because the international participants had little experience of eating the target food and had a narrower range of expectations about the taste of the target food. Looking at the gender rows of **CrTaS** and **CrTaC** in Figure 4.9, the female participants felt the stronger taste of steamed rice and the weaker taste of curry and rice when visually presented with curry and rice compared to the male participants. In other words, the female participants' taste perceptions were less modulated by visual stimuli, which is the same trend as that of the noodle conditions.

These results again suggest that the taste perceptions of the female participants were less modulated by visual stimuli. However, looking at the gender row of **CrTyC** in Figure 4.10, they felt like they were eating curry and rice better than male participants in the **Cr** condition. As in the noodle conditions, the results suggest that the participants perceived a change in the type of food even if they did not perceive a change in the food's taste.

Like in the noodle conditions, we have to note that the generalizability of our findings is limited due to the small number of participants despite these findings.

4.3.4 Correlation between Noodle and Rice Conditions

We here discuss the correlation between the noodle and rice conditions. Comparing the two modulation conditions, the VAS scores in the **Cr** conditions (see Figure 4.7, **CrTaC** and **CrTyC**) are lower than those in the **Fn** conditions (see Figure 4.5, **FnTaF** and **FnTyF**). We believe that this is because the steamed rice was near tasteless whereas the somen noodles were not. Gustatory manipulation becomes more difficult when the gap in taste between the original and target types of food is larger. Comparing the numbers of Table 4.1, **FnTaS** and **FnTaF** and Table 4.4, **CrTaS** and **CrTaC**, those in the rice conditions are smaller. We believe that this is not because the gustatory manipulation is more stable but because it is weaker in the rice conditions.

The Pearson's product moment correlation coefficient between the noodle and rice conditions indicates a weak relationship on the VAS scores of the perceived taste of the modulated food (**FnTaF** and **CrTaC**) ($r = 0.39$), and that of the recognized type of food (**FnTyF** and **CrTyC**) ($r = 0.33$). These results suggest that the participants who scored high in the **Fn** condition also scored high in the **Cr** condition and that the multisensory flavor perceptions of some participants are more strongly affected by visual modulation than others.

4.3.5 Overall Discussion

We here summarize the above discussions and answer the questions that arose in the previous user study. For the first question (whether and how the gustatory manipulation persists while eating the modulated food), since the significant differences for persistency were not confirmed in almost all the cases, we can say that the gustatory manipulation persists to some extent for a longer period of time. Meanwhile, there were individual differences in the tendency for persistent changes in visual modulation. In particular, there were many individual differences in the perception of food types during the visual modulation, and there were similar number of participants who gradually felt more strongly that they were eating the original food, those who felt that they were eating the food presented by the visual modulation, and those who did not change. In other words, their tendency for vision-induced multisensory flavor perception change persistent exists in three groups: those in which the intensity of the multisensory flavor perception change gradually increased with each biting session, those in which it gradually decreased, and those in which it did not fluctuate.

For the second question (whether and how not seeing the original food before seeing the modulated food affects the strength of gustatory manipulation), this experiment revealed that the gustatory manipulation occurred and the strengths of the perceived taste and the confidence of the recognized food are similar whether or not participants saw the original food before seeing the modulated food under both the noodle and rice conditions. We hypothesize that the intensity of the taste change would be enhanced when participants were not presented with the actual food they were eating, but these results did not support our hypothesis.

For the third question (whether and how the results depend on the participants' nationality and gender), we confirmed that the strength of gustatory manipulation varies depending on nationality and gender. In the present experiment, non-Japanese and male participants felt the stronger gustatory manipulation compared to Japanese and female participants, respectively. We speculate that the more they are familiar with the original and target types of food, the weaker gustatory manipulation they will feel due to more accurate expectations of the food experience. The experiments focusing on nationality and gender showed these trends, but note that the number of participants was small. An additional large-scale study with many participants is needed to examine the effects of cultural differences such as nationality and gender on vision-induced multisensory flavor perception.

4.4 Limitations

We believe that the reported user study is valuable in general, however, it also has several limitations. Here, we discuss some of these limitations and future directions.

4.4.1 Food Types

We tested only two original types of food (steamed rice and somen noodles) and two target types of food (curry and rice and fried noodles). We chose rice and noodles because of the popularity in East Asia. Somen noodles and steamed rice are both widely available and known for weak tastes thus good as original types of food. Fried noodles and curry and rice were selected due also to the popularity. In addition, they often contain 'forbidden' ingredients such as pork. We believe that the selection of these food types is reasonable considering that we target to support people with dietary restrictions in the future. We would like to conduct follow-up studies on the relationship between the strength of gustatory manipulation and the visual or gustatory similarities between the original and target food types.

4.4.2 Participants

Even though we increased the geographical, ethnic, and gender variations among the participants compared to those in the previous user study, the low variations in age and the background (most were graduate students) are possible limitations of this study. For example, the small number of participants might have had a significant impact on the results. The large average age difference between males and females is problematic in comparing the effects of gender, and the bias in the country of origin of international participants is problematic in comparing the differences between Japanese and international participants. However, it should not be ignored that despite these problems, there was a clear tendency for differences in food experience to affect multisensory flavor perception that promises interesting follow-up studies in the future.

We would like to conduct further studies using different combinations with greater variation of food types and sample population in the future.

4.4.3 Experimental Protocol

In our user study, participants knew what they were actually eating regardless of the visual modulation conditions even though they did not see the original food without the HMD. One way to avoid unwanted bias would be to mix the opposite modulation conditions (e.g., fried noodles as the original food and somen noodles as the target food). However, such additional conditions will significantly increase the cost of food preparation.

4.4.4 Image Translation Quality

The image quality of visual modulation needs to be improved significantly. Twitter images used for training were generally taken close to the food with chopsticks or a spoon whereas the participants saw the food from slightly farther positions. Visual modulation sometimes failed because of this difference. Besides, the current system converts the entire image into the target food so we needed to run the experiment in a texture-less environment. Another problem was that our system's output was coarse and more or less the same for one type of food (e.g., fried noodles) whereas there are actually many variations within a single type of food. Because of this, the modulated food looked very different from participants' expectations sometimes. In the future, we plan to improve the neural network by introducing many new features, such as food region extraction, supporting a larger number of food types and high resolution images [124, 125].

4.4.5 System Latency

The system latency has been improved from 400 ms in Chap. 3 to about 150 ms in the current system. However, it is still non-negligible. Even though no-one reported relevant problems such as motion sickness or nausea, the latency must be further shortened for a practical use.

4.4.6 Difficulty of Eating

Many participants reported the difficulty of eating while wearing the HMD. This was mainly due to the horizontal offset between the participants' eyes and the cameras for the VST experience. Because of this, they thought their mouth position was about 10 cm away from their actual position. In the future, we would like to use a custom-designed VST HMD with a smaller horizontal parallax and a wider opening around the mouth.

4.5 Conclusions

In this chapter, we have reported a user study on the effectiveness of our GAN-based gustatory manipulation system primarily from the perspectives of persistency, nationality, and gender differences.

Our experimental results revealed that vision-induced gustatory manipulation is persistent in many participants. Their persistent multisensory flavor perception changes are divided into three groups: those in which the intensity of the multisensory flavor perception change gradually increased, those in which it gradually decreased, and those in which it did not fluctuate, each with similar number of participants. Our results also revealed that those participants (e.g. males and international participants) who are less familiar with the original and target types of food feel stronger gustatory manipulation.

We believe that our research has provided a deeper understanding and insights into GAN-based gustatory manipulation. We hope many researchers will be encouraged to conduct follow-up studies. In the future, we will improve the image translation quality, further reduce the system latency, and minimize the horizontal parallax. Then we will conduct follow-up studies using a wider variety of foods for a larger number of participants from a more diverse demographic background.

5 Effects on multisensory flavor perception when the surrounding environment is changed to a virtual environment

5.1 A Video See-through Food Overlay System for Eating Experience in the Metaverse

We have reported a gustatory manipulation interface that changes the perception of taste and type of food through visual modulation using GAN in Chap. 3 and Chap. 4. The results of these experiments revealed that a visual modulation technique that alters the appearance of food could change the taste and type of food, and these changes persist. On the other hand, visual modulation using HMDs can change not only the appearance of food but also the surrounding environment into the VE. The VE gives the user the sense of being there or presence through a VR display such as a HMD [128]. The sense of being in VEs is called presence [128]. In particular, social VEs such as VRChat [†] and Mozilla Hubs [‡] enable users to communicate with other users using avatars, which are virtual bodies in the virtual world. Social VEs, also known as the metaverse [129], are attracting a great deal of commercial and cultural attention. The metaverse has great potential to change our lives. For example, Korsgaard *et al.* reported the elderly preferred eating with friends in the VE to eating alone [20]. However, wearing the HMD makes it difficult to eat because users cannnot see the RE. Many current users of the metaverse eat by either viewing the food through the small gap around the nose or by removing the HMD. These eating experiences significantly prevent users from maintaining the presence of the VEs. We believe that eating in the metaverse should be easy and enjoyable.

Meiselman *et al.* found that the ratings differed depending on the surrounding environment [15]. Dionisio *et al.* used a CAVE for visual presentation to provide a highly immersive meal in VEs [76]. Eating while wearing an HMD requires the presentation of real food images, and it is necessary to fuse VST with VE images. Korsgaard *et al.* switched the transparency of the VST and VE images depending on the direction of the head [18]. Switching between VST and VE video has been a problem that reduces presence because participants are aware of the RE. Several studies have also evaluated superimposing edited VST video onto the VE using color detection and chroma key composition [19, 20]. These studies require the hardware that unifies desks and tableware and limit the user experiences. Further, these studies aimed for stand-alone VR applications and could not be used with existing VR applications. In other words, users need to be in a specific VE to dine with a high degree of presence.

These problems also reduce the degree of freedom and applicability when studying gustatory manipulation using visual modulation techniques that alter the surrounding environment. For example, researchers who want to conduct dietary research include doctors and cooks, not necessarily VR experts. They currently cannot use the completed VEs or Metaverse that have been uploaded and must create their own VEs from scratch when they conduct gustatory research in VEs.

In this chapter, we have developed Ukemochi [††] (named after the Japanese god of food) that enables users to eat while maintaining a high presence by superimposing only the food region on an existing VR application. Ukemochi uses the overlay function of OpenVR API [‡‡], which can be used simultaneously with general-purpose VR applications such as VRChat and Mozilla Hubs using OpenVR API. Ukemochi also detects and tracks the food region, requiring less preparation than the chroma key method. In this chapter, we compared and reported three eating methods in the metaverse (VE only, VST displayed, and Ukemochi).

We also investigated the effects of changes in the surrounding environment on multisensory flavor perception by examining the effects of whether the images around food were of the RE (VST displayed) or VE (Ukemochi) or whether the food was not visible (VE only), on the taste and flavor of the food.

The major contributions of this chapter are as follows:

- We propose Ukemochi that enables users to eat while maintaining the high presence of existing VEs.
- We show that superimposing only the food segmentation image on the VE gives a high presence.
- We report that superimposing food images on VE improves ease of eating and that even superimposing only the segmented food regions, which reduces the visibility of the RE, maintains the similar level of ease of eating.
- We also report that changing the appearance of the food's surroundings had no effect on taste and flavor and no statistical difference was found.

5.2 System

Ukemochi enables users to see and eat real food while maintaining a high sense of presence in VR applications. In this chapter, we use the front camera of an HTC VIVE to obtain real food images. As shown in Fig. 5.1, Ukemochi consists of client and server modules. The client was developed using Unity 2018.4.f1 and has three functions: obtain real food images using a VIVE front camera, send the images to the server module, and overlay real food images on VR applications. The server is implemented in Python 3.8.8 and Flask 1.1.2, and it produces food segmentation images from the real food image via object tracking. (a) The client obtains an RGB image (612×460) from the VIVE front camera and (b) sends it to

[†] VRChat, `https://hello.vrchat.com/`, last accessed March 17, 2023.

[‡] Mozilla Hubs, `https://hubs.mozilla.com/`, last accessed March 17, 2023.

[††] Ukemochi, `https://signs0302.github.io/ukemochi`, last accessed March 17, 2023.

[‡‡] SteamVR API, `https://store.steampowered.com/app/250820/SteamVR`, last accessed March 17, 2023.

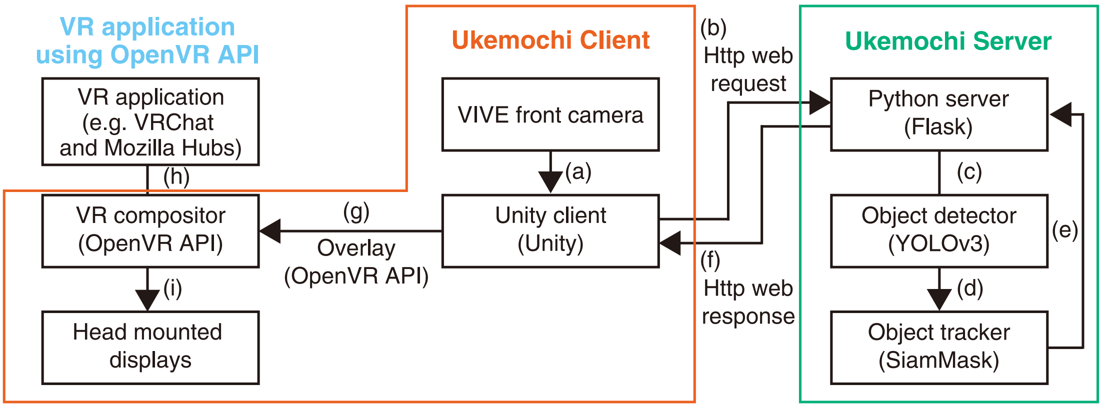


Figure 5.1: Flow of Ukemochi system, which consists of two modules – an Ukemochi client and server module.

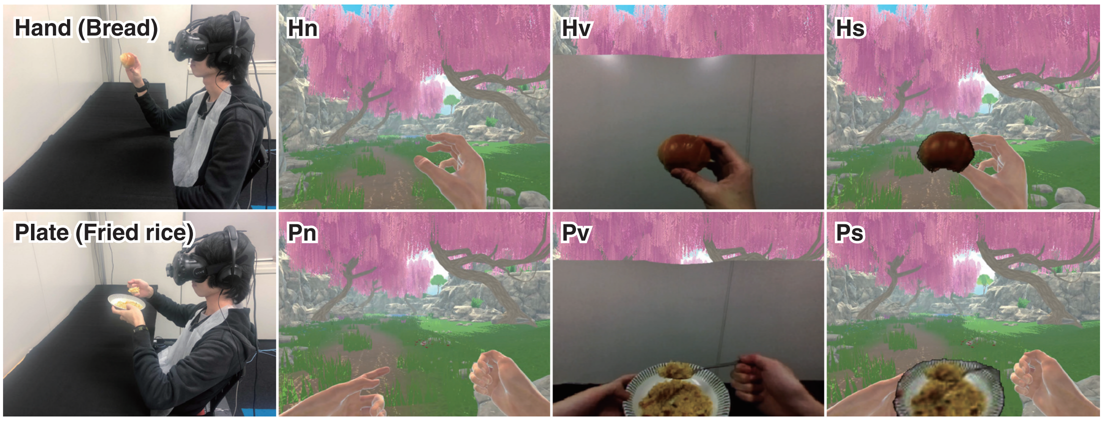


Figure 5.2: Food scenes using Ukemochi in VEs. Top) Conditions for eating bread by hand. Bottom) Conditions for eating fried rice on a plate. **Hand**, **Plate**) The real-world environment for each meal condition. **Hn**, **Pn**) **None** condition without the real image. **Hv**, **Pv**) **VST** condition with the superimposed VIVE front camera image. **Hs**, **Ps**) **Seg** condition with only the food region cropped.

the server through an HTTP web request. (c) The server creates an RGB image (306×230) by removing the distorted regions from the received image from the client (d, e) and generates a food segmentation image by using an object tracker (SiamMask [85] trained with VOT-2018 [130]) that is initialized by a result of an object detector (YOLOv3 [131] trained with UECFood100 [132]). (f) The food segmentation image is sent from the server to the client. (g) The client calculates the position of the food segmentation image for the Unity coordinate system and overlays it. (h) The food segmentation image is then superimposed on the running VE image using the overlay function of OpenVR API. (i) Thus, the user can see the VE with the real food image while wearing an HMD (see Fig.5.2 (**Hs**, **Ps**)).

In the experiments, we run Ukemochi on a desktop computer with an AMD Ryzen Threadripper PRO 3975WX 3.50 GHz CPU, a 64GB memory, and two NVIDIA RTX 3060ti GPUs. The frame rate was about 21 [fps] and the video delay was about 150 [ms]. We applied AR Timewarping [126] to reduce an apparent latency caused by the rotation of the HMD.

5.3 Experiment

5.3.1 Overview

In this experiment, we investigated whether changing the visual presentation method of the food displayed in the VE changes the sense of presence, food evaluation, and ease of eating. The experiment setup consisted of an HTC VIVE, a Leap Motion, noise-canceling headphones, and a PS4 controller. Ukemochi was developed not to require desk preparation in a uniform color. However, preliminary experiments suggested that Ukemochi was in that case unstable and could have reduced the participants' presence. Thus, we conducted the experiment in a room with a black desk and white walls to ensure the system's stable operation as shown in Fig.5.2 (**Hand**, **Plate**). Participants ate two kinds of food (bread and fried rice) after experiencing the VE for 3 minutes and answered a questionnaire. The VE and avatar were used as shown in Fig.5.2 (**Hn**, **Pn**). They were instructed to explore the VE freely for 3 minutes using a PS4 controller. They could move around by the left stick and change the direction by the right stick. The Leap Motion captured a hand motion to animate the avatar's hands synchronized with the user's. We let them adjust the avatar's hands to be the same size as their own.

We compared three visual conditions to display food images. In **None** condition, participants could see the food through the gap between the HMD and their noses (see Fig.5.2 **Hn**, **Pn**). In **VST** condition, the participants could see the raw image of the front camera including the food region (see Fig.5.2 **Hv**, **Pv**). In **Seg** condition, the participants could see the food segmentation image superimposed on the VE (see Fig.5.2 **Hs**, **Ps**). Participants performed an eating task with all the three visual conditions in a counterbalanced order based on a Latin square design.

In addition, the participants experienced two meal conditions. They primarily differ in the style of eating; eating the food directly in hand (bread) and on the plate with a spoon (fried rice). One butter roll (Topvalu best price butter rolls 6 pcs, 27 grams each, Aeon group) was used per condition for the **Hand** condition. One hundred grams of fried rice was used for the **Plate** condition (Authentic stir-fried rice®, Nichirei foods Inc.) which was cooked just before the experiment. The participants ate food in a fixed position in the VE in all conditions (see Fig. 5.2). They experienced one visual condition per day for three days. They also experienced all meal conditions in each of the three days. The order of the meal conditions was balanced and fixed for each participant. In addition, participants were verbally confirmed before the experiment that they were neither extremely hungry nor highly full.

5.3.2 Hypotheses

We set the following hypotheses.

H1 **Seg** condition with food segmentation results in the highest presence and better Igroup Presence Questionnaire (IPQ) [133] scores.

H2 **Seg** condition with food segmentation results in the highest evaluation of the taste of food.

H3 **Seg** condition with food segmentation results in the highest evaluation of the appearance of food.

H4 The ease of eating food is the highest in **VST** condition but not significantly different than in **Seg** condition because seeing the non-food regions is expected to have little effect.

5.3.3 Procedure

The participants answered a questionnaire on their age and VR experience before the experiment. First, participants were briefed on the operation of the PS4 controller. Next, they were fitted with a disposable apron and the HMD. Then, they adjusted the size of the avatar's hands. The following steps were then performed under each visual condition.

1. Participants explore the VE for 3 minutes using the controller.
2. They move to the eating spot in the VE.
3. They grab the food of the plate with their hands.
4. The experimenter starts Ukemochi and presents the participants with three visual conditions.
5. Participants begin to eat their food after the start signal.
6. After the participants feel that they have finished eating the food, they will give the end signal aloud.
7. They remove the HMD and complete the questionnaire.
8. They drink a cup of water and then repeat Steps (1) to (7) under different meal conditions.

Step (2) was introduced to make the visual experience consistent among the participants. The water drinking action in Step (8) was done to reduce the influence of the previous condition. After experiencing all visual conditions, participants completed a post-questionnaire that included forced rankings and free comments.

5.3.4 Questionnaire

We asked the participants how much experience they had with AR and VR (How much experience have you had with AR and VR?). The participants answered in three levels (I have no experience at all, I have little experience, and I have experienced it many times). The questionnaire included questions to evaluate the

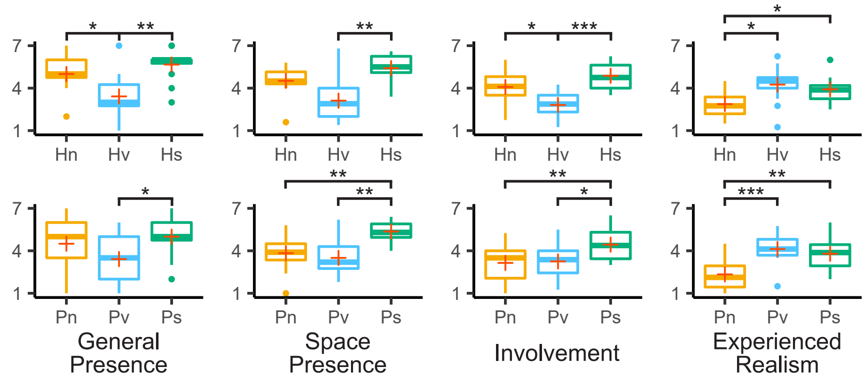


Figure 5.3: IPQ results. Top) Results for **Hand** condition. Bottom) Results of **Plate** condition. The red crosses indicate the mean values, and the dots indicate the outliers. [1 (strongly disagree) to 7 (strongly agree)].

sense of presence, eating itself, and ease of eating. The IPQ questionnaire [133] was used to evaluate presence. In this experiment, four of the subitems of the IPQ (*Pres*: General Presence; *SP*: Space Presence; *Inv*: Involvement; and *Real*: Experienced Realism) were rated using a seven-point Likert scale [1 (strongly disagree) – 7 (strongly agree)]. *Pres, SP, Inv*, and *Real* consisted of one, five, four, and five items, respectively. In addition, in this experiment, the sum of the mean values of each evaluation item (*Total*: Total Presence) was calculated as the evaluation index. *Taste and flavor* and *Appearance of food* were assessed using the VAS method based on Hannan-Jones and Capra's questionnaire [134]. In addition, the VAS was used to evaluate the *Ease of eating*. After all the conditions were completed, participants gave a ranking to four questions related to the visual conditions: "*Taste*: Taste of the food", "*Easy*: Ease of eating the food", "*Feel*: Feeling of the experience", and "*Presence*: Immersion or presence in the VR space."

5.4 Results and Discussions

5.4.1 Results

Twelve (nine males and three females; mean age = 25.0 years; standard deviation = 7.02) participated in the study. In response to their experience with AR and VR applications, eleven participants answered "I have experienced it many times," and one participant answered "I have no experience at all." The main results are shown in Figs. 5.3 and 5.4 and Table 5.1. We performed a post hoc analysis with the Holm–Bonferroni correction for the results in Figs. 5.3 and 5.4. We also performed a Friedman's test followed by a post hoc analysis with the Least significant difference for the results in Table 5.1. Significant differences are indicated with symbols (*** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$).

5.4.2 Summary of main findings

The main findings are as follows.

1) In terms of *Total* results of the IPQ scores, participants who experienced **Seg** condition (**Hs**, **Ps**) using Ukemochi perceived the highest presence. In addition, *Feel* and *Presence* results showed that the method in **Seg** condition gave the highest sense of presence, indicating that the proposed method was preferred as an experience. These results show that the proposed method of displaying only the food segmentation image gives a higher presence, supporting [**H1**]. The *Pres*, *Sp*, and *Inv* results of the IPQ scores show that **Seg** condition gives the highest presence. Especially in *Sp* and *Inv* results, a significant difference was found between the **Plate** conditions **Pn** and **Ps**. Several participants commented that "I found it very difficult to eat under the **Pn** condition and concentrate on eating." In other words, the eating difficulty in the **Pn** condition may have

Table 5.1: Results of forced ranking. The values in the left three columns are the average of the ranking (1st to 3rd). The Freedman column shows the significant difference of the Friedman's test. The results in the right three columns are the significant differences between each condition using the Least significant difference.

	None	VST	Seg	Freedman	None & VST	None & Seg	VST & Seg
Taste	2.67	1.67	1.67	*	*	*	
Easy	2.92	1.33	1.75	***	***	*	
Feel	2.50	2.33	1.17	**		**	*
Presence	2.33	2.50	1.17	**		***	***

reduced presence. The *Real* result was the highest in **VST** condition (**Hv**, **Pv**). This result is an indicator of the realism of the VE experience, and we believe that the VST experience was perceived as realistic. On the other hand, there was no significant difference between **VST** condition and **Seg** condition. In other words, the **Seg** condition provided the similar level of realism as **VST** condition.

2) There was no significant difference in the Taste and flavor results of the VAS score between the conditions. The Taste result of the ordered answers showed significant differences between **None** condition and the other conditions. These results indicate that the food segmentation slightly improves the taste of the food, which partially supports [**H2**]. In addition, the rankings of **VST** and **Seg** conditions were higher than that of **None** condition, suggesting that improving the visibility of food improves the evaluation of taste in the VE.

3) The *Appearance of food* result of VAS score showed a significant difference between **None** condition and the other conditions. This result indicates that the improved visibility of the food improves the evaluation of the appearance of the food, which partially supports [**H3**]. We speculated that **Seg** condition would result in the highest ratings, but the results showed that **VST** condition had the highest ratings. The participants said, "The resolution of the superimposed food segmentation image in **Seg** condition was low and felt uncomfortable." and "The food's appearance became worse when the food segmentation failed." Therefore, it is considered that the resolution and accuracy of the food region segmentation need to be improved.

4) The *Ease of eating* result of VAS score showed a significant difference only between the **Hn** and **Hv** conditions. This result suggests that the visibility of the food improves the ease of eating. However, there was little difference in the **Plate** condition. In the **Plate** condition, it was necessary to use a plate and a spoon to eat while wearing the HMD, and the eating action was challenging in all conditions. Therefore, we believe that this is a reason for no significant differences. On the other hand, we may find significant differences between **None** condition and other conditions as the number of participants increases (**Hn** to **Hs**, **Pn** to **Pv**, and **Pn** to **Ps**: $p = 0.063$). In addition, the *Easy* result of the ordered answer showed significant differences between **None** condition and the other conditions, and **VST** condition was rated higher than **Seg** condition. We believe that these

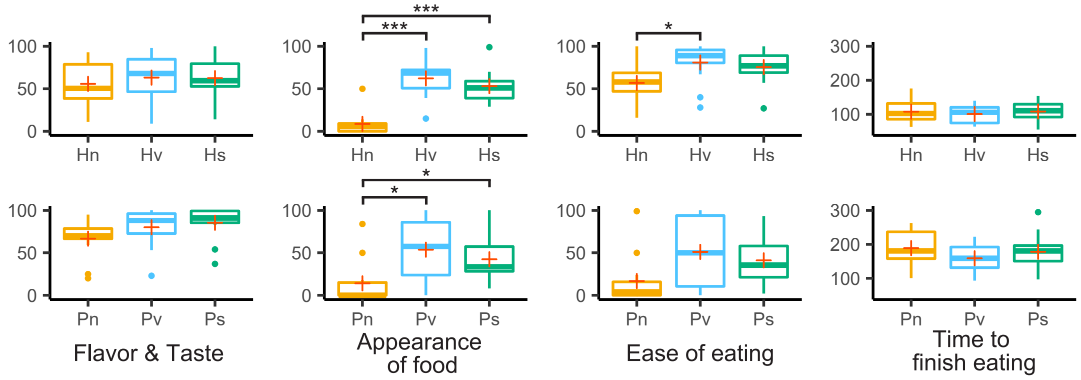


Figure 5.4: VAS results and the time to finish eating result. Top) Results for **Hand** condition. Bottom) Results of **Plate** condition. The red crosses indicate the mean values, and the dots indicate the outliers. The three graphs from the left show VAS results [0 (poor) to 100 (excellent)]. The rightmost graph shows the time to finish eating (s).

overall results partially support **[H4]**. No significant difference was detected in the results of "Time to finish eating". We believe that this is due to the short time to finish eating.

5.4.3 General discussions

Efficacy of food visualisation

According to the experimental results, it was found that our method of displaying only the food segmentation image using Ukemochi provided a high presence. In addition, we found that our proposed method provides a similar level of ease of eating as **VST** condition using the VST images. We expect that Ukemochi can be used to maintain a high presence in many VEs. On the other hand, the evaluation of eating ease was lower than our expectation. One of the reasons for this was that we could not present images around the mouth due to the limited vertical viewing angle of the HMD. It is expected that the ease of eating will be improved with an HMD with a wider vertical field of view [135]. It may also be improved by increasing the resolution and accuracy of the food segmentation image. No significant difference was observed in the evaluation of the taste and flavor of the food. Several participants commented that “I felt a sense of discomfort between the atmosphere of the VE and the realistic images” and “The food looked terrible because of the low resolution.” Therefore, it is expected that the deliciousness of the food can be improved by increasing the resolution of the images, generating images that match the atmosphere of the VEs [136], and replacing realistic images with VR objects [137].

Effects of visual rendering quality

Most of the participants considered the low resolution of the VST images to be a problem. In particular, in the **Plate** condition, several participants commented that “I could not see how many rice grains were left on the plate because of the low resolution.” In the **Hand** condition, if the bread becomes too small, the food region segmentation fails, and the participant’s arm or the entire raw image is presented as food. These results suggest the necessity of detecting the amount of remaining food. Participants responded in terms of the VEs experience that “I enjoyed the view of the VE when in **Seg** condition.” In addition, participants reported in the **Pn** condition that “I needed to concentrate on eating, which detracted from the VE experience.” Moreover, the *Feel* result of ordered answers was the highest in **Seg** condition. It suggested a need to present the food and VE images in an appropriate ratio to improve the VE experience. Some participants commented that “I wanted someone to eat with them to improve the quality of food in the VE.” We plan to improve Ukemochi so that users can eat together in VEs in the future.

5.5 Limitations

As a system capability limitation, Ukemochi often fails to detect out-of-distribution food samples over UECFood100, which is a training dataset for the object detector. To solve this problem, we need to build a wide variety of dataset that includes foods from all over the world. Additionally, we cannot evaluate a performance of the object tracker for our purpose because there are no food annotated videos. To conduct fair experiments, we should build an evaluation dataset. Participants reported that the food did not look tasty because of the low resolution of the segmented food images. The low resolution of the input image and the frame-by-frame inference required a lower resolution. The video delay and low frame rate also degraded the user experience. Despite these limitations, Ukemochi requires high computer specifications to operate. To solve these problems, we should reduce the amount of computation. The images of food displayed by Ukemochi are not visible to others and cannot be viewed in 3D because they are 2D images. We should add the ability to share food images through server-to-server communication and the ability to replace 2D images with 3D virtual objects. Limitations of the experimental design include the small number of food items used in the experiment, the small number of participants, and the skewed gender distribution. We should experiment on a messy desk to evaluate Ukemochi in daily use. We should consider that tracking and food segmentation errors could have affected the experimental results. The participants had difficulty in answering the questionnaire because it was difficult to determine whether the VST images superimposed on the VE were real or virtual. We believe that this problem can be solved by using an augmented reality-compatible questionnaire [138].

5.6 Conclusion

In this chapter, we developed and investigated the effectiveness of Ukemochi that detects and tracks the food segmentation image and overlays on the VEs. We found that displaying only segmented food images overlaid on the VE gives the user a higher presence compared to displaying raw video frames or no food images in the VE. We also found that the food segmentation images had a similar level of ease of eating compared to the raw images. On the other hand, the taste and appearance of the food were not affected favorably due presumably to the problems of low resolution and detection accuracy, which we plan to improve in the future. We plan to follow up with features such as transforming food segment images to match the atmosphere of the VE, replacing objects in the VR, and allowing users to eat together. Despite the limitations described above, we believe that Ukemochi will maintain a high presence while eating in VEs and will facilitate the study of eating in the rapidly developing metaverse.

6 Presentation of visual information near the mouth

6.1 HMD with a wide downward field of view

We have reported gustatory manipulation interface using cross-modal effects of visual modulation with VST HMD in Chap. 3 and Chap. 4. This gustatory manipulation interface generated a cross-modal effect by changing the appearance of food into different foods using GAN. HMDs present a food images generated by GAN with the use of high-resolution flat displays such as LCDs and eyepiece optics. However, the FoV of a typical HMD for VR is limited to approx. 90–110 [deg] in the diagonal direction and approx. 70–90 [deg] in the vertical direction, which is narrower than that of humans, especially in the vertical direction. The vertical human FoV is approx. 120–135 [deg], and the downward FoV below the horizontal line of sight (approx. 70–80 [deg]) is larger than the upward FoV (approx. 50–55 [deg]). Conventional HMDs typically cover only approx. 50% of the downward FoV of humans. Therefore, it is impossible to modulate the food's appearance in the mouth's vicinity, which may inhibit gustatory manipulation. We also have reported the problem of reduced ease of eating due to the lack of visibility in the mouth's vicinity in Chap. 3, Chap. 4, and Chap. 5. Much effort has been made to increase the FoV of an HMD, such as the use of free-form optical elements and multiple LC panels [139] or curved displays and curved microlens arrays [88]. However, as far as we have searched only a few studies focused on the vertical FoV.

In this chapter, we developed a novel HMD with a pair of additional display units to increase the downward FoV to address the aforementioned problems. We investigated whether the increase in the downward FoV improves the amount of cross-modal effect and ease of eating. In addition, it was clear from the participants' comments that the smell and food texture changed somen noodles to fried noodles while they were eating the somen noodles that had been changed into the appearance of fried noodles in Chap. 3 and Chap. 4. Therefore, we also investigated whether visual modulation can change the smell and food texture.

The major contributions of this chapter are as follows:

- we developed a novel HMD with additional display units for a wide downward FoV.
- We report that in the experiment with the downward FoV expansion, significant differences were not found in the ease of eating while wearing an HMD and the amount of gustatory manipulation of the cross-modal effect using visual modulation.
- We have demonstrated that gustatory manipulation by visual modulation can change the smell and food texture from the original food to the intended food.

6.2 HMD that expands the downward FoV

6.2.1 System configuration of VR part

A schematic diagram and an actual prototype of the proposed device are shown in Figs. 6.1 and 6.2, respectively. We removed the exterior of the HTC VIVE and placed two sets of an LCD module (Sharp LS029B3SX02; size, 2.9 [inch]; display area, 51.84×51.84 [mm]; resolution, 1440×1440; frame rates, 120 [fps]) and a Fresnel lens (length, 35 [mm]; width, 38 [mm]; focal length, 40 [mm]) at an angle of 20 [deg] under each eyepiece with a three-dimensional (3D) printed housing. The eye relief and the distance from the lens to the LCD were both 28 [mm]. This is shorter than the lens's focal length, but the image is observed clearly. The distance between the user's eyes and the HTC VIVE eyepiece was adjusted by approx. 13 [mm] to the longest to provide the space for the additional display units. We used Unity 3D to develop the software for the proposed system. Two virtual cameras were used to present images on the LCDs, which followed

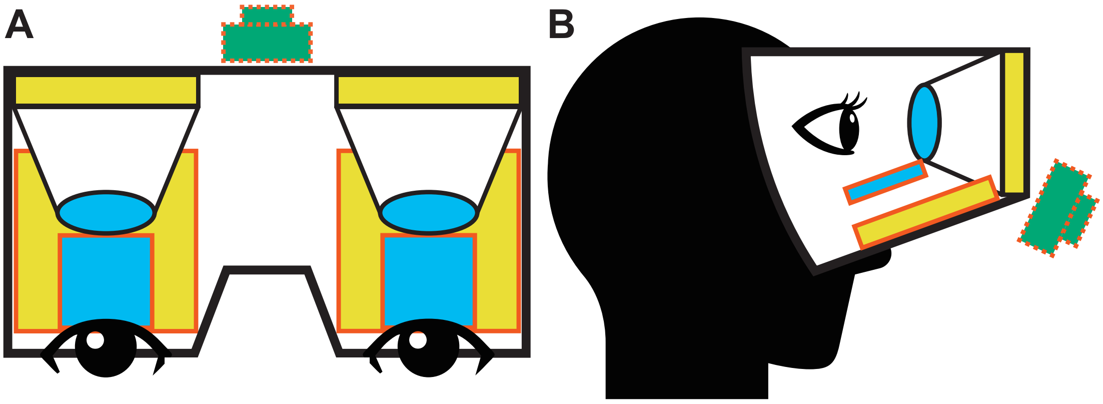


Figure 6.1: Schematic of the proposed HMD. A) Top view, B) Side view: yellow, display units; light blue, eyepiece; light green, fish eye camera; black lines, exterior of the HTC VIVE; orange lines, newly added parts.

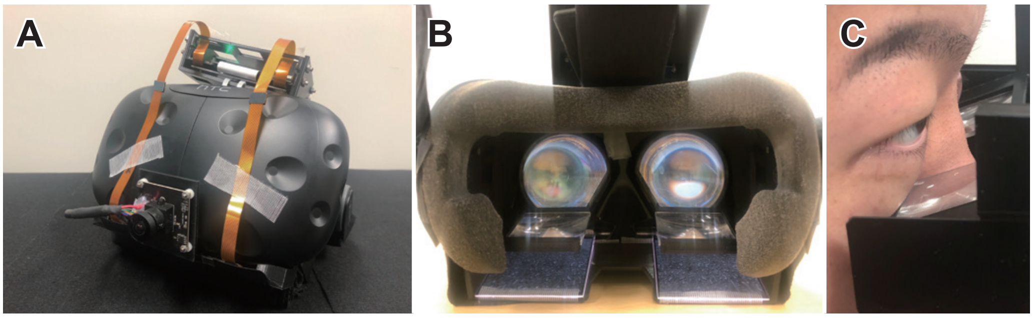


Figure 6.2: Prototype HMD. A) Front view, B) Back view, C) Bottom display placement relative to the face.

the main camera of SteamVR. The virtual cameras' initial intrinsic and extrinsic parameters were given by considering the physical dimensions of the display units. The edited view frustum corrected the image from the virtual sub-camera, and then the distortion caused by the Fresnel lens was manually corrected by mesh deformation (Figure 6.3). We checked the stereo visibility with several individuals and fixed the IPD of the VIVE and that of the lower optics to 69.4 [mm] and 70 [mm], respectively, as the optimal common distance.

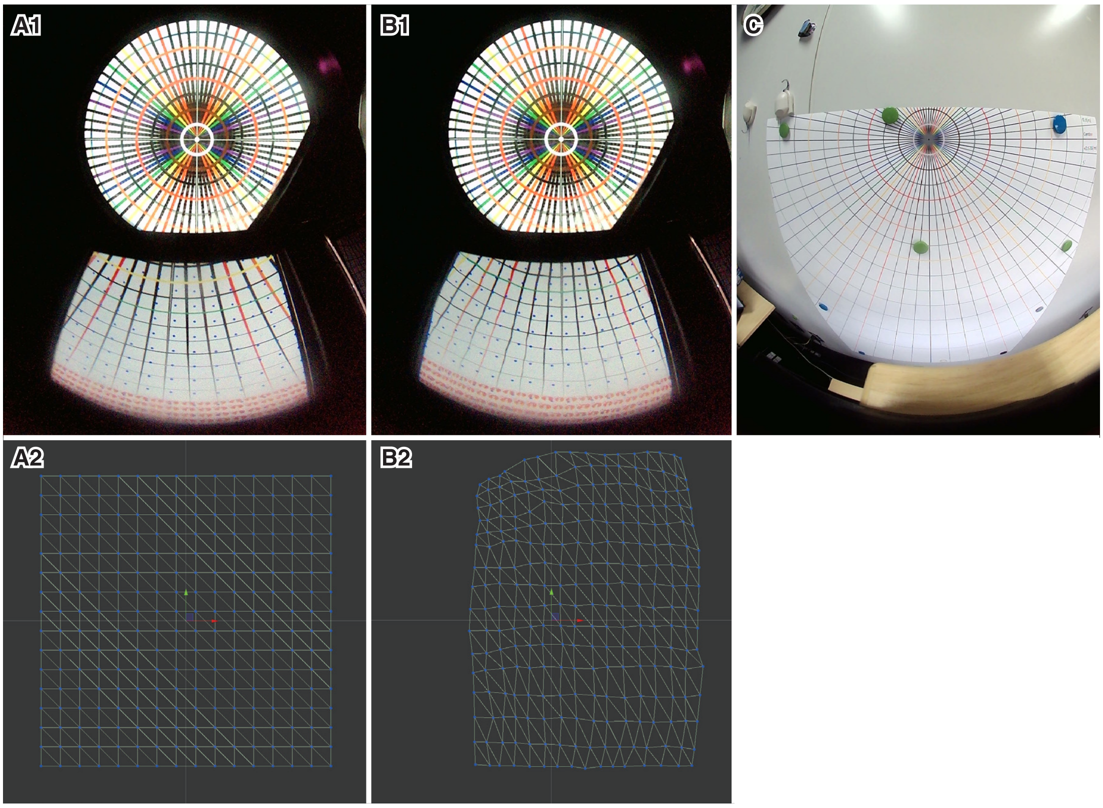


Figure 6.3: Images displayed on the HMD and the mesh shape before and after mesh deformation. A1) Display image before mesh deformation captured by the fisheye camera. A2) Mesh shape before deformation. B1) Display image after mesh deformation captured by the fisheye camera. B2) Mesh shape after deformation. C) Correct pattern placed in front of the fisheye camera, each line represents a viewing angle in 5 [deg] increments.

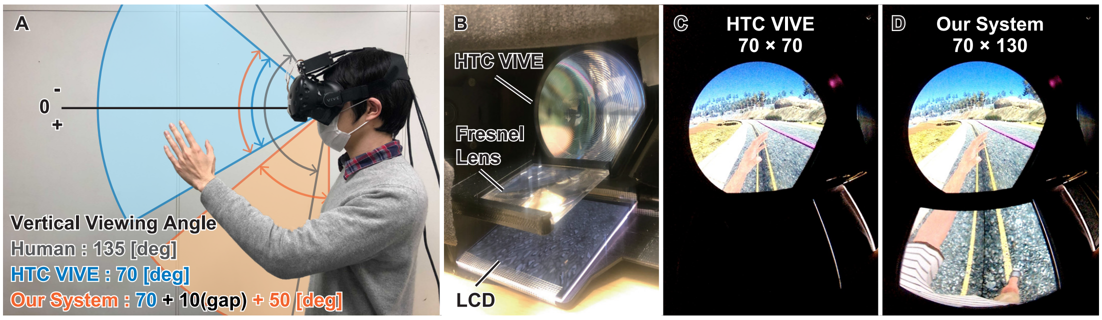


Figure 6.4: Our HMD with a wide downward FoV. A) Side view. Blue: vertical FoV of the HTC VIVE. Part of the lower FoV is missing owing to the Liquid Crystal Displays (LCDs) housing. Orange: increased vertical FoV. Gray: vertical FoV of a typical human. B) The added LCD and Fresnel lens. C) Original FoV of the HTC VIVE (approx. 70×70 [deg]). The user's hand is visible but not the lower limbs and body. D) FoV of the prototype (approx. 70×130 [deg]). The user's lower limbs and body and the ground are all visible at the same time.

6.2.2 FoV of the prototype HMD of VR part

Figure 6.4 C and D show how the VE is presented with the prototype HMD. The prototype is designed to have a downward FoV of 90 [deg], which is sufficiently large to fully cover the downward FoV of human vision (approx. 70 [deg]). The actual FoVs measured by a pre-calibrated equidistant projection fisheye camera (GS-15WDCM-1.5MM; resolution, 1920×1080; FoV, 180 [deg]) are shown in Table 6.1. Note that the FoV of the HTC VIVE officially announced by HTC is approximately 110 [deg] in the horizontal, vertical, and diagonal directions because of its circular viewport, and the actual FoV reported by iNFINITE is approximately 89 [deg] in the horizontal and vertical directions [§§]. As can be seen from the table, the implemented prototype increases the downward FoV by approx. 60 (10 + 50) [deg]. The slight decrease in the FoV of HTC VIVE from the official specification may be due to the larger distance between the eyepiece of HTC VIVE and the user's eye (an increase of approx. 13 [mm]) in our setting.

[§§] iNFINITE, `https://www.infinite.cz/projects/HMD-tester-virtual-reality-headset-database-utility`, last accessed March 17, 2023.

The diagonal FoV of the original HTC VIVE is approx. 75 [deg], but the horizontal FoV is slightly smaller owing to the missing part of the Fresnel lens near the nose. The downward FoV of the original HTC VIVE is also reduced by approx. 10 [deg] owing to the 3D printed housing. A dedicated fused lens can prevent the FoV reduction in the future [‖].

6.2.3 System configuration of VST part

We develop an HMD with an increased downward FoV of VST area that can present visual information in the downward FoV (Figure 6.5). In order to provide visual modulation to foods near the mouth, a fisheye camera (ELP-SUSB1080P01-L170; resolution, 1080×1920; FoV, 170 [deg]; frame rate, 50[fps]) was added to the front of our HMD prototype, tilted 25 [deg] downward from the front.

The actual FoVs of VST measured by a pre-calibrated equidistant projection fisheye camera (GS-15WDCM-1.5MM; resolution, 1920×1080; FoV, 180 [deg]) are shown in Table 6.2.

6.2.4 Visual modulation function using machine learning

For changes in food appearance, we used a system based on StarGAN [73] from a previous study [118]. The system consists of a client created with Unity 2019.4.39f1 and a server created with the Python framework Flask. First, the

[‖] Panasonic, `https://channel.panasonic.com/contents/19737/`, last accessed March 17, 2023.

Table 6.1: FoVs measured by an equidistant projection fisheye camera. The diagonal FoV of the original HTC VIVE is approx. 75 [deg].

FoV	Horizontal	Vertical (upper + gap + lower)
Original HTC VIVE	∼70 [deg]	∼70 (40 + 0 + 30) [deg]
Our prototype	∼70 [deg]	∼130 (40 + 10 + 80) [deg]
Difference	±0 [deg]	+60 (0 + 10 + 50) [deg]

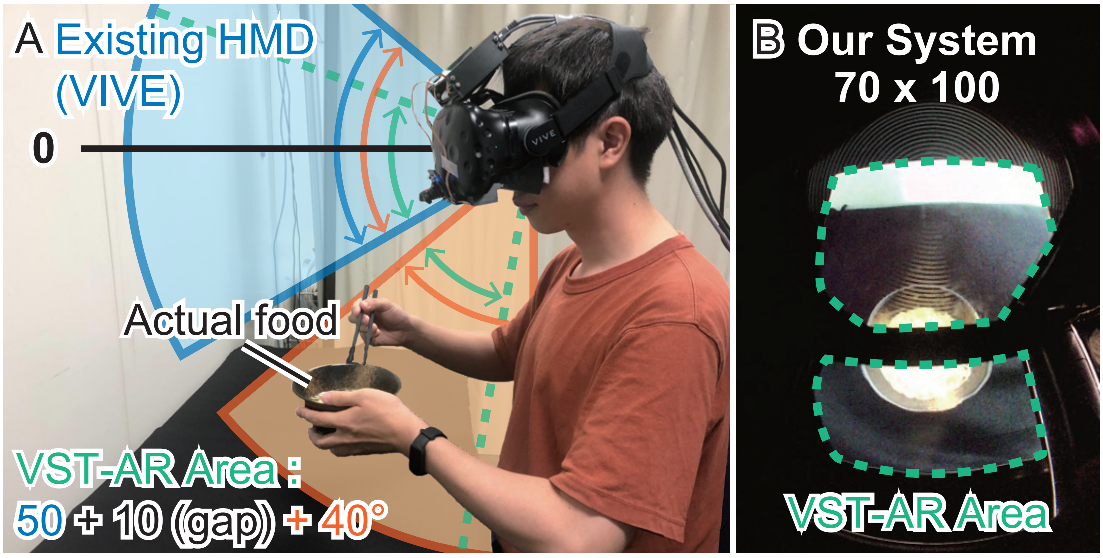


Figure 6.5: Our VST-HMD with a wide downward FoV. A) Side view. Blue: vertical FoV of the HTC VIVE. Part of the lower FoV is missing owing to the LCDs housing. Orange: increased vertical FoV. Green: vertical FoV of a VST–AR area. B) FoV of the prototype (approx. 70×100 [deg]). The upper and lower parts of the actual food are visible at the same time.

server acquires the images from the fisheye camera added to the front of the HMD. Next, the distortion of the fisheye camera was corrected using OpenCV, and a visually modulated food image was generated using StarGAN. The resolution of the fisheye camera was limited to 480×640 to reduce the delay. The final food image resolution was 256×341 due to reducing the fisheye camera distortion correction and resizing for image generation.

The measured delay was approximately 200 [ms]. Finally, the client displayed the generated food image 1 [m] away from the user's eyes, tilted 25 [deg] downward like a fisheye camera.

Table 6.2: This table is the FoV of the VST area, measured with an equidistant projection fisheye camera. FoV is shown as a range because a portion of the VST image is not displayed in the added downward viewing area.

FoV	Horizontal	Vertical (upper + gap + lower)
Original HTC VIVE	∼70 [deg]	∼50 (20 + 0 + 30) [deg]
Our prototype	∼50–70 [deg]	∼100 (20 + 10 + 55–70) [deg]
Difference	±0 [deg]	+35–50 (0 + 10 + 25–40) [deg]

6.3 Experiment

6.3.1 Overview

In this experiment, we investigated whether the increased visibility of food near the mouth due to an enlarged downward FoV improves the amount of effect of gustatory manipulation by visual modulation and the ease of eating food when wearing the HMD. We also investigated the cross-modal effect of sensing the smell and texture of food presented by visual modulation, which was reported in a previous study [118].

As the actual food, we used cups of somen noodles, the same food as in the previous study [118, 140]. Only the soup powder was used, and no condiments were used. The experiment was conducted according to a factorial design with 2×2. The independent variables were the downward FoV (*ON*: the additional display units were used; *OFF*: the units were *not* used) and the food appearance (Sn: original food *without visual modulation* (somen noodles), Fn: trancereted food *with visual modulation* (fired noodles)). Thus, we have a total of four conditions (OffSn: Off Somen noodles, OnSn: On Somen noodles, OffFn: Off Fried noodles, OnFn: On Fried noodles).

The appearance of the food under each condition is shown in Figs. 6.6. All variables were within-participant, and participants performed the tasks in a counterbalanced order under all conditions based on the latin square design.

This experiment was evaluated using a questionnaire for sensory evaluation and quantitative evaluation based on head angle and elapsed time while eating.

6.3.2 Procedure

The experiment was conducted in a quiet room in our laboratory. The room is furnished with a black desk and chair facing the white wall, with red marks drawn on them so that they appear to be in front of them when they sit down (Figure 6.7 (B)). Participants were asked to confirm that they were not moderately hungry and full. After sitting at a desk, they were told about the experiment in which they were to eat the somen noodles under each condition, and they confirmed the questionnaire. The experiment was then conducted in the following order.

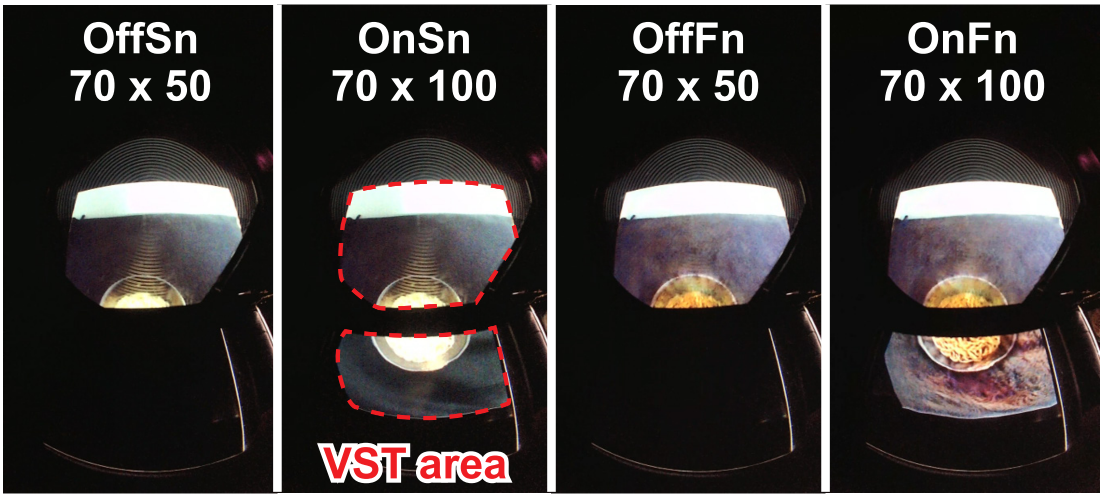


Figure 6.6: Viewing angle and video output for each condition. OffSn, OffFn) horizontal: approx. 70 [deg], vertical: approx. 50 [deg]. OnSn, OnFn) horizontal: approx. 70 [deg], vertical: approx. 100 [deg] (maximum viewing angle displayed))

1. Participants drink water.
2. They wear HMDs with increased downward FoV.
3. They hold the bowl and chopsticks in their hands for 10 seconds in each condition and look at the somen noodles in the bowl.
4. They place the bowl and chopsticks on the table and look so that the red markings on the wall are visible in front of them.
5. They are told to begin the task.
6. They eat more than two bites of somen noodles using chopsticks.
7. After looking at the red markings on the wall so that they are visible in front, they announce the end of the task.
8. After they remove the HMD, they fill out a questionnaire.
9. Repeat steps 1–8 for each condition.

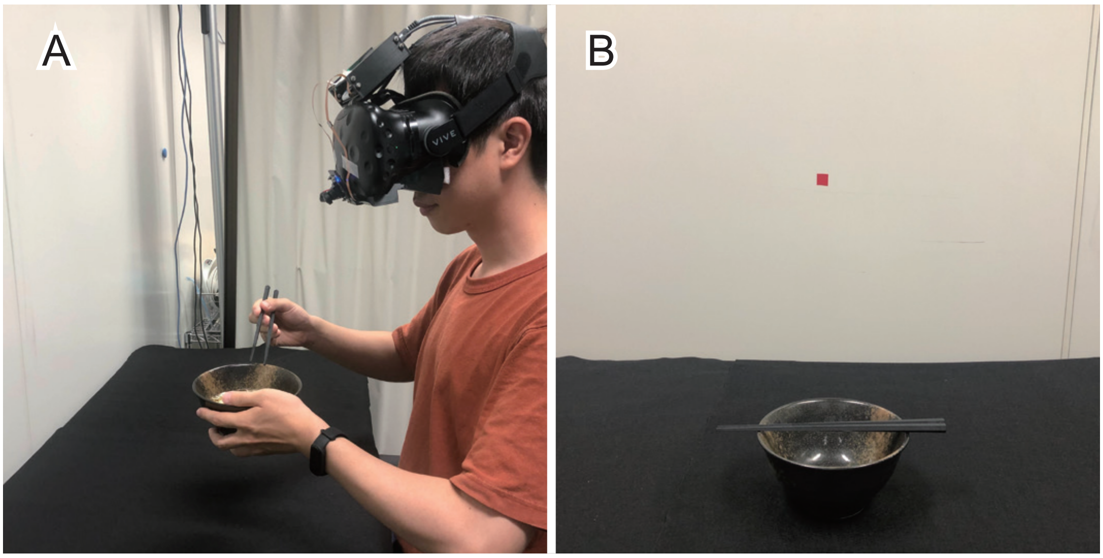


Figure 6.7: Diagram of the experiment. A: Side view, B: Red markings on the desk and wall used in the experiment, bowls, and chopsticks.

Step 1 was performed to wash out the food remaining in the mouth in each condition. Step 3 was performed to check the appearance of the food in each condition. The movement to see the red marks in steps 4 and 7 was performed to unify the initial angles when measuring the pitch angle. The elapsed time of the experiment was measured in the interval from steps 5 to 7. Food was shown to participants only while wearing the HMD not to affect the experimental results.

6.3.3 Questionnaire

The questionnaire consisted of eleven questions measuring the taste, type, appearance, smell, texture, and ease of eating to investigate the effect of the downward FoV expansion on multisensory flavor perception.

The detailed questions are shown below:

Q1. It tasted like somen noodles.

Q2. It tasted like fried noodles.

Q3. It felt like I was eating somen noodles.

Q4. It felt like I was eating fried noodles.

Q5. It felt the appearance of food was somen noodles.

Q6. It felt the appearance of food was fried noodles.

Q7. It felt the smell of somen noodles.

Q8. It felt the smell of fried noodles.

Q9. It felt the food texture of somen noodles.

Q10. It felt the food texture of fried noodles.

Q11. It was easy to eat with the VR goggle (HMD).

Q1 to **Q**4 were used to measure perceived taste and recognized type of food similar to previous studies [118, 141] to investigate the effect of the downward FoV expansion on flavor perception. **Q**5 and **Q**6 were used to measure whether visually presented food images appeared to be the food we intended. **Q**7 to **Q**10 were used to measure whether the cross-modal effects of visual changes affected olfaction and tactile. **Q**11 was used to measure the ease of eating. All questions were asked in random order on a scale of 101 using the VAS [0 (strongly disagree) to 100 (strongly agree)].

6.3.4 Hypotheses

In this experiment, we set the following hypotheses.

H1 Increasing the downward FoV improves the amount of cross-modal effect on gustation by improving the visibility of the food after visual modulation, which in turn found an interaction in the results of **Q**1, **Q**2, **Q**3, and **Q**4.

H2 Increasing the downward FoV improves the visibility of the food after visual modulation, which is felt in the appearance of the visually presented food, which in turn found an interaction in the results of **Q**5 and **Q**6.

H3 Cross-modal effects with visual modulation evoke the smell of the presented food, which in turn decreases **Q**7 scores and increases **Q**8 scores in the **F**n condition.

H4 Cross-modal effects with visual modulation evoke the food texture of the presented food, which in turn decreases **Q**9 scores and increases **Q**10 scores in the **F**n condition.

H5 Increasing the downward FoV improves the amount of cross-modal effect on olfaction, which in turn found an interaction in the results of **Q**7 and **Q**8.

H6 Increasing the downward FoV improves the amount of the cross-modal effect on tactile, which in turn found an interaction in the results of **Q**9 and **Q**10.

H7 Increasing the downward FoV improves the ease of eating because the food is visible near the mouth, which in turn increases **Q**11 scores and decreases **T**ime scores in the **O**n condition.

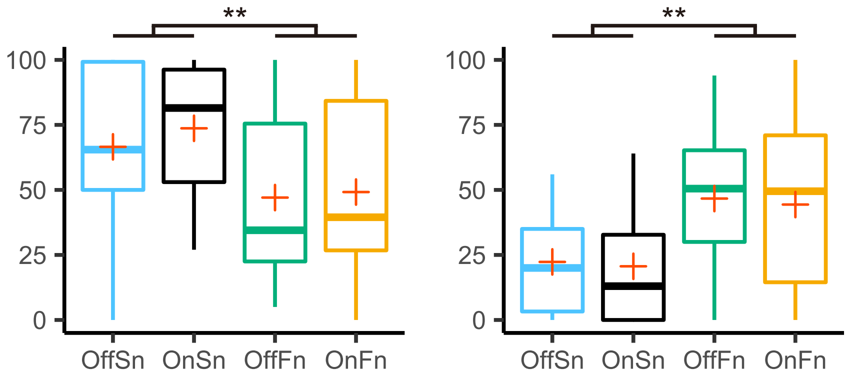
H8 Increasing the downward FoV improves the ease of eating because the food is visible near the mouth, which in turn upward the head pitch angle and reduces variance in the **O**n condition.

6.3.5 Participants

Participants were recruited through a campus mailing list and announcements. In accordance with the ethical review committee of the author's institution, informed consent was obtained from each of the participants after the study was fully explained to them. Each participant was paid equivalent to about 10 USD. Six teen (eight males and eight females; mean age = 26.1 years; standard deviation = 8.89) participated in the study.

6.4 Results

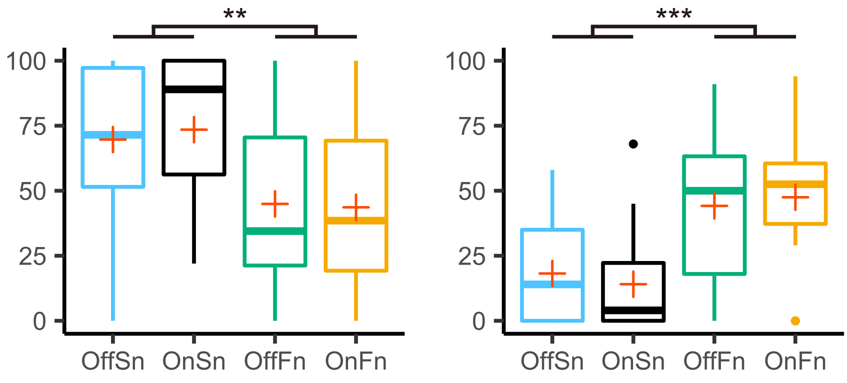
The results are shown in Table 6.3 and Figs. 6.8, 6.9, 6.10, 6.11, 6.12, 6.13, and 6.14. The red crosses in each figure indicate average values. We performed two-



It tasted like
Somen noodles [Q1]

It tasted like
Fried noodles [Q2]

Figure 6.8: The perceived taste of food results (Q1, Q2) [0 (strongly disagree) to 100 (strongly agree)].



It felt like I was eating
Somen noodles [Q3]

It felt like I was eating
Fried noodles [Q4]

Figure 6.9: The recognized type of food results (Q3, Q4) [0 (strongly disagree) to 100 (strongly agree)].

way ANOVA (the additional LCD units: ON vs. OFF and the food appearance: before modulation (Sn) vs. after modulation (Fn)). Since all data were not normally distributed, we employed the aligned rank transform for hypothesis testing [127]. Significant differences are indicated with symbols (*** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$).

Table 6.3 shows the results of the questionnaire, as well as the average meal time [s] (Time) and the average head pitch angle [deg] (Average, Variance, and

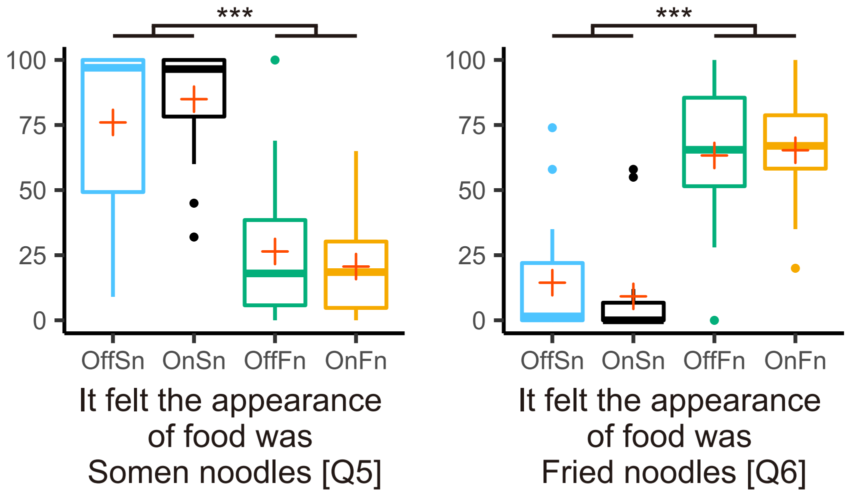


Figure 6.10: The perceived appearance of food results (Q5, Q6) [0 (strongly disagree) to 100 (strongly agree)].

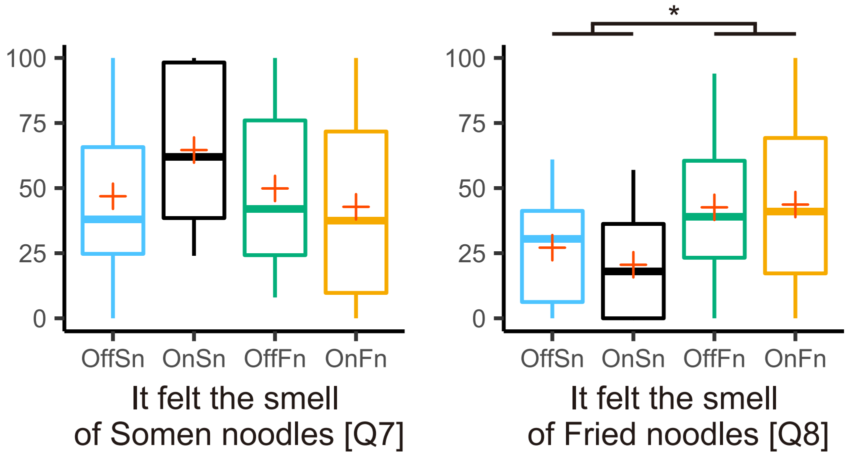


Figure 6.11: The perceived smell of food results (Q7, Q8) [0 (strongly disagree) to 100 (strongly agree)].

Maximum).

Figure 6.8 shows the results of *Q1* and *Q2*, which investigated the perceived taste of food. Two-way ANOVA revealed a significant main effect of the food appearance (Sn vs. Fn) on *Q1* ($p < 0.01$) and *Q2* ($p < 0.01$). No significant main effect of LCDs (ON vs. OFF) and interaction were found.

Figure 6.9 shows the results of *Q3* and *Q4*, which investigated the recognized

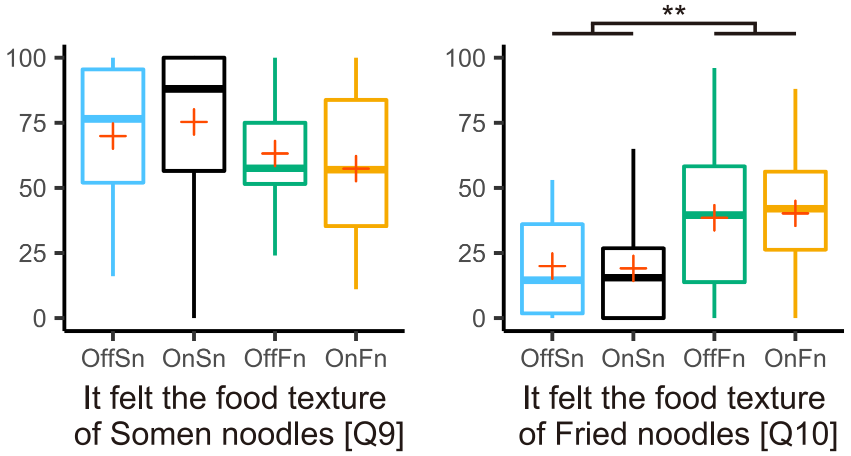


Figure 6.12: The perceived food texture results (Q9, Q10) [0 (strongly disagree) to 100 (strongly agree)].

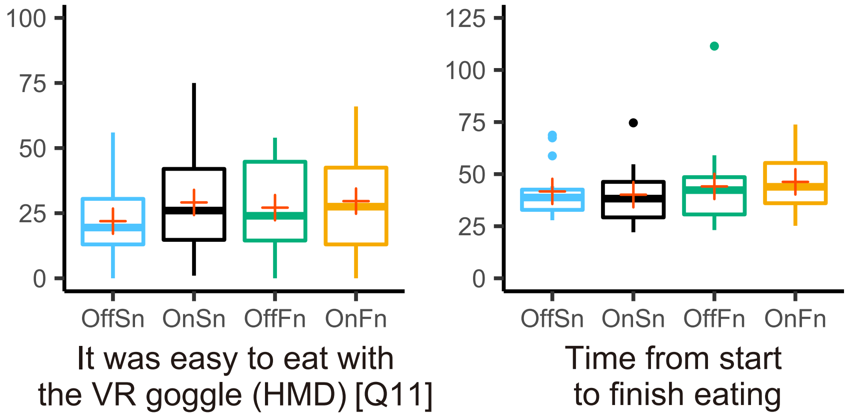


Figure 6.13: Left) The ease of eating results (Q11) [0 (strongly disagree) to 100 (strongly agree)]. Right) The time results while eating [s]

type of food. Two-way ANOVA revealed a significant main effect of the food appearance (Sn vs. Fn) on *Q3* ($p < 0.01$) and *Q4* ($p < 0.001$). No significant main effect of LCDs (ON vs. OFF) and interaction were found.

Figure 6.10 shows the results of *Q5* and *Q6*, which investigated the perceived appearance of food. Two-way ANOVA revealed a significant main effect of the food appearance (Sn vs. Fn) on *Q5* ($p < 0.001$) and *Q6* ($p < 0.001$). No significant main effect of LCDs (ON vs. OFF) and interaction were found.

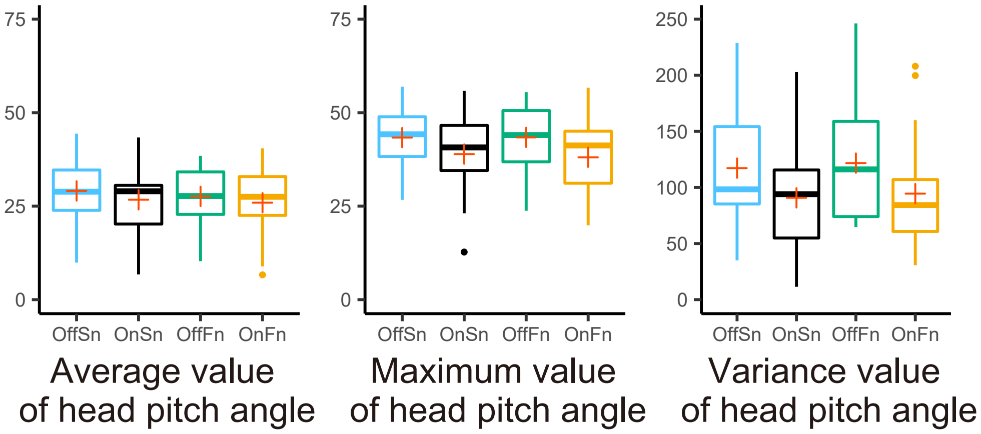


Figure 6.14: The head pitch angle while eating results [deg]. Left) Maximum value. Middle) Maximum value. Right) Variance value. The head pitch angle is 0 [deg] when facing front and +90 [deg] when fully facing down.

Figure 6.11 shows the results of *Q7* and *Q8*, which investigated the perceived smell of food. Two-way ANOVA revealed a significant main effect of the food appearance (Sn vs. Fn) on *Q8* ($p < 0.05$). No significant main effect of LCDs (ON vs. OFF) and interaction were found.

Figure 6.12 shows the results of *Q9* and *Q10*, which investigated the perceived food texture. Two-way ANOVA revealed a significant main effect of the food appearance (Sn vs. Fn) on *Q10* ($p < 0.01$). No significant main effect of LCDs (ON vs. OFF) and interaction were found.

Figure 6.13 shows the result of *Q11* and the time results while eating, which investigated the ease of eating. Also, Figure 6.14 shows the result of head pitch angle, which investigated average, maximum, and variance. Two-way ANOVA did not find any significant differences in the *Q11*, time, and head pitch angle (Average, Maximum, Variance).

Table 6.3: Experimental results. Each value represents the mean and the standard deviation. Items with a significant difference are in bold.

		OffSn	OnSn	OffFn	OnFn
Taste	**Somen noodles**[Q1]	**66.56±29.43**	**73.69±26.04**	**47.06±33.12**	**49.19±35.89**
	Fried noodles[Q2]	**22.31±20.27**	**20.63±23.38**	**46.69±29.59**	**44.38±32.31**
Type	**Somen noodles**[Q3]	**69.69±28.86**	**73.5±29.65**	**44.94±31.74**	**43.63±34.88**
	Fried noodles[Q4]	**18.19±19.68**	**14.06±20.85**	**44.19±30.06**	**47.50±28.31**
Appearance	**Somen noodles**[Q5]	**76.00±32.72**	**84.94±22.24**	**26.44±28.26**	**20.63±19.38**
	Fried noodles[Q6]	**14.44±23.46**	**9.19±18.85**	**63.31±26.49**	**65.31±21.49**
Smell	Somen noodles[Q7]	46.88±32.28	64.63±30.08	49.88±31.05	42.81±35.08
	Fried noodles[Q8]	**27.13±20.47**	**20.56±20.23**	**42.63±30.56**	**43.69±34.23**
Texture	Somen noodles[Q9]	69.88±29.00	75.31±29.51	63.19±22.56	57.38±31.40
	Fried noodles[Q10]	**19.94±19.39**	**19.06±22.28**	**38.50±29.62**	**40.19±27.11**
Easy	Ease of eating[Q11]	21.94±15.86	29.13±19.49	27.13±17.61	29.63±20.09
Time	Time	41.71±12.51	40.13±13.90	44.11±20.94	46.28±14.49
Head	Average	29.07±8.41	26.72±9.29	27.61±8.43	25.93±10.33
	Maximum	43.34±7.91	38.92±11.11	43.39±9.08	38.07±10.60
	Variance	117.25±56.48	90.77±53.52	121.72±51.38	94.49±53.48

6.4.1 Discussion

The main findings are as follows.

1) The results from **Q**1 to **Q**4 show that taste and food type perception are altered by changes in the appearance of food (Sn vs. Fn). These results suggest that, with previous studies [118, 140], gustatory manipulation by visual modulation is possible even in experiments using HMDs with an increased downward FoV.

However, there was no significant main effect of LCDs (ON vs. OFF) and interaction was found, which did not support [H1]. These results suggest that even if the increased downward FoV improves the visibility near the mouth, the amount of effect of gustatory manipulation using the cross-modal effect may not change. A possible explanation for this is that the participants only sometimes saw the food until they brought it near their mouths and put it in their mouths. Therefore, it is possible that the amount of effect of the gustatory manipulation did not change because participants did not see the food after modulation, even though the visibility near the mouth was improved.

Another possible explanation for this is that the cross-modal effect of visual modulation functioned maximally in both conditions and that the amount of effect of gustatory manipulation did not change. To investigate these hypotheses, we believe it is necessary to add eye trackers to HMDs with an increased downward FoV, measure eye movement while eating, and follow up to see how much the downward FoV is used.

2) The results of **Q**5 and **Q**6 show that the changes in food appearance (Sn and Fn) occur as we intended. These results indicate that visual modulation also occurs in experiments using HMDs with an increased downward FoV.

However, there was no significant main effect of LCDs (ON vs. OFF) and interaction was found, which did not support [H2]. These results suggest that increased downward FoV does not tend to make the meal appear even more like the intended food. A possible explanation is that there was a non-negligible gap between the front and bottom displays of around 10 [deg] owing to the 3D printed housing, which prevented a unified borderless FoV. Although none of the participants reported that this gap had a negative impact, and it could even be regarded as glasses, reducing the gap could improve the overall experience. We will develop an improved system using dedicated eyepieces such as a fused lens ‖.

3) The results of **Q**7 show that was no significant main effect of food appearance (Sn vs. Fn) was found. Meanwhile, **Q**8 show that changes in the appearance of food (Sn vs. Fn) indicate that participants smelled fried noodles in the **F**n condition. These results partially support [H3]. In short, the cross-modal effect of visual modulation on olfaction does not decrease the smell of actual food, but increases the smell of visually presented food.

These are interesting result, as it differs from the cross-modal effect on gustation, which tends to decrease the taste/type of the actual food and increase the taste/type of the presented food. Figure 6.11 (Left: **Q**9) has a lower score for **O**ffSn compared to **O**nSn, indicating a tendency not to perceive the smell as somen noodles. Similarly, **O**ffSn scores for Figure 6.10 (Left: **Q**7) were lower than those for **O**nSn, indicating that the respondents tended not to perceive the appearance of the food as somen noodles. In the limited downward FoV of the **O**ffSn condition, participants did not perceive the food they saw as somen noodles, so it is possible that they did not perceive the odor as somen noodles. Note that the **Q**7 statistics are the main effect of food appearance ($p = 0.24$) and the interaction ($p = 0.14$).

4) The results of **Q**9 show that was no significant main effect of food appearance (Sn vs. Fn) was found. Meanwhile, **Q**10 show that changes in the appearance of food (Sn vs. Fn) indicate that participants smelled fried noodles in the **F**n condition. These results partially support [H4]. In short, the cross-modal effect of visual modulation on tactile does not decrease the food texture of actual food, but increases the food texture of visually presented food.

These result on tactile are similar to the result on olfaction, as it differs from the cross-modal effect on gustation, which tends to decrease the taste/type of the actual food and increase the taste/type of the presented food. Meanwhile, the results of **Q**9 shows a marginal trend toward significance on food appearance ($p < 0.10$) was found. Changes in food texture may be more likely to occur than changes in the smell of food. Futurer studies should be to determine how the cross-modal effects of vision on olfactory and tactile change the actual smell and food texture.

5) The results from **Q**7 to **Q**10 show that were no significant main effect of

LCDs (ON vs. OFF) and interaction was found, which did not support [H5] and [H6]. These results suggest that increasing the downward FoV to improve the visibility of the mouth suggests that the magnitude of the cross-modal effect does not change, similar to the results from **Q**1 to **Q**4.

6) The results of **Q**11 and **T**ime show that were no significant main effect of LCDs (ON vs. OFF) and interaction was found, which did not support [H7]. The results of **H**ead pitch angle also show that were no significant main effect of LCDs (ON vs. OFF) and interaction was found, which did not support [H8]. Contrary to expectations, these results suggest that an increase in the downward visual field may not affect the ease of eating or the pitch angle of the head.

A possible explanation for the no change in participants' feeling the ease of eating [**Q**4] might be not looking at the food near the mouth when their eating. Participants may have no problem seeing that they can grasp the food with chopsticks on the front display and may not need a downward FoV. We believe it is necessary to measure the movement of the participant's point of view with eye-tracking technology that supports a downward FoV.

Another possible explanation for this might be that participants were unfamiliar with the AR experience using HMD. We reported on a preliminary experiment involving a small group of our lab members specializing in VR research and familiar with the AR experience [142]. Eight (five males and three females; mean age = 22.9 years; standard deviation = 0.83) participated in the preliminary experiment. The preliminary experiment was evaluated with a limited number of questionnaires **Q**1 to **Q**4, **Q**11; the **H**ead pitch angle was not measured.

Figure 6.15 shows a comparison between the results of this experiment and a preliminary experiment on the ease of eating. The results in the preliminary experiment on ease of eating show a significant main effect of LCDs (ON vs. OFF) was found ($p < 0.001$). These results suggest that ease of eating with HMD varied depending on the participants' demographics. It should be noted that familiarity with the AR experience was not measured and that the participants in the preliminary experiment may have had a bias to improve the results.

We believe that the lack of significant differences in the head pitch angle and time results was due to the inability to set the measurement interval correctly. The participants determined the timing of the end of the measurement section

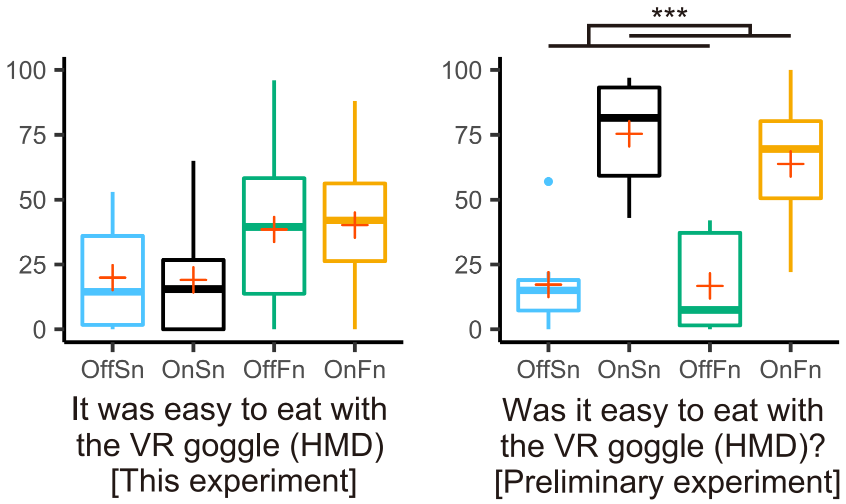


Figure 6.15: The ease of eating results [0 (strongly disagree) to 100 (strongly agree)]. (Left) This experiment. Right) Preliminary experiment [142]

because we could not visually confirm whether the participants had finished their food. They were also free to move their head angle not only when eating the food but also when they were not looking at the food until they tasted it and completed the measurement. Future experiments are needed to measure the head angle only during the interval between bringing the food to their mouths and taking it into their mouths. The results in the preliminary experiment on maximum and Variance of head angle show a marginal trend toward significance on LCDs ($p < 0.10$) was found. These results show that the ease of eating that an enlarged downward FoV improves the ease of eating cannot be ruled out and requires further study.

6.5 Conclusion

In this chapter, we developed a VST-HMD with an increased downward FoV. We also investigated whether the increased visibility of food near the mouth due to the expanded downward FoV improves the ease of eating and the amount of cross-modal effects caused by visual modulation. The experimental results did not find any statistical difference in the amount of cross-modal effect and in improving the ease of eating, despite the visual modulation to the mouth area. Although the increase in downward FoV was not effective in manipulating multisensory flavor perception, it did allow us to create a system to investigate the influence of the appearance of the food near the mouth on perception. We also investigated whether visual modulation can change the smell and food texture from the original food to the intended food. As a result of the experiment, we presented fried noodles' smell and food texture even when eating somen noodles. In the future, we plan to add an eye tracker to the increased downward Fov HMD and conduct a follow-up study to measure how users view the food from the desk when they bring it to their mouths and to clarify how they perceive objects near their mouths. We also investigated whether visual modulation can change the smell and food texture from the original food to the intended food. As a result of the experiment, we presented fried noodles' smell and food texture even when eating somen noodles. These results suggest that gustatory manipulation by visual modulation is flexible and applicable.

7 The effect of the avatar in the virtual environment changes on flavor perception

7.1 Presentation of the avatar without virtual mirrors

This chapter presents the impact of displaying an avatar in the VE in the downward FoV. There may seem to be no connection between the display of avatars in the downward FoV and the study of multisensory flavor perception. However, we believe that there are two major advantages of using avatar displays in gustatory research: (1) the avatar display improves the likelihood of visual information, and (2) changes in the avatar's physical characteristics may change flavor perception.

We have reported the effects of visual modulation on the appearance of food (Chap. 3 and Chap. 4) and changes in the surrounding environmental appearance (Chap. 5) on multisensory flavor perception. These flavor perceptions result from the multisensory integration of stimuli from vision and gustation. Such multisensory integration is known to occur more frequently in peri-personal space, which is the space within the reach of the human [21, 22]. It is believed that multisensory integration occurs in humans when multiple spatially and temporally consistent senses are inferred to be the common cause. When sensory stimuli are considered to emanate from different sources, they are inferred to be two perceptions each rather than one integrated perception. To be inferred as a common cause, a sensory stimulus must be plausible for a prior distribution based on previous experience. However, it isn't easy to see one's body, which is always visible in daily life, without virtual mirrors installed in the VE in studies using HMDs.

This is because it is difficult to see the self-body other than the hand due to the limitation of the downward FoV. We believe that the avatar display problem could be solved by using the HMD with an increased downward FoV, as created in Chap. 6.

The downward FoV plays an important role in recognizing our own bodies. We believe that the downward FoV is particularly important for SoE (SoSL, SoA, and SoBO) [23]. For example, hand–torso connectivity is one of the most important factors in evoking SoA and SoBO [105, 106]. However, the shoulders and upper arms are typically outside the FoV of conventional HMDs. Moreover, the downward FoV is important for awareness of the ground and the visibility of the user avatar's feet, which are important when walking [113, 115]. A wide downward FoV could also improve SoSL by increasing feet visibility and increase the clarity regarding where the self-avatar is standing. In addition, limiting the downward FoV negatively affects user behavior, such as increasing the downward head pitch angle [113], walking time, and the number of obstacle contacts [114]. When the downward FoV is limited in the real environment, users tend to look down carefully when walking down a staircase [115]. When the downward FoV is limited in the VE, a user may look at his or her feet more often out of fear. Increasing the downward FoV may also improve the immersion (presence and SoE) in the VR experience by improving the user's unnatural head movements/angles and bringing them closer to the head movements in the real environment.

Thus, we believe that the avatar body display in the downward FoV improves the likelihood of visual information. We expect that the display of avatars will improve the effect of gustatory manipulation through multisensory flavor perception.

Furthermore, research is being conducted to change the behavior of users and improve their abilities by changing the avatars' appearance [30, 31]. For example, Kilteni *et al.* reported that in an experiment of playing drums in the VE, the avatar of a musician with afro hair beat the drums more rhythmically than the avatar in a business suit [30]. Banakou *et al.* reported that using an avatar that looked more like Einstein improved test performance than using an avatar that looked more like the experimental participant [31]. This effect of changes in the avatar's body affecting the psychological state and behavior of humans is known as the Proteus effect and is being actively studied [32]. Based on these research reports, we hypothesized that changes in the avatar's body might affect multisensory flavor perception. For example, the user may perceive a meal as sweeter when the avatar he selects is a woman, the delicate taste of food when it is a chef, or the luxury of a meal when it is a millionaire. Thus, changing the avatar's appearance can alter flavor perception by changing its vision. However, as mentioned above, the limited downward FoV limits the display of all but the avatar's hands. We need to investigate the effectiveness of the avatar's body display for the downward FoV as a preliminary step to investigate the effect of the avatar's body change on multisensory flavor perception.

In this chapter, We investigated whether the increase in the downward FoV improves presence, SoSL, SoA, SoBO, cybersickness, and head movement patterns in the VR experience using the HMD with increased downward FoV developed in Chap. 6. The major contributions of this chapter are as follows:

- We demonstrated that the HMD developed in Chap. 6 with an increased downward FoV could display avatars and change the appearance. It could potentially be used to investigate its effects on multisensory flavor perception.
- We elucidate the trend that the enlargement of the downward FoV improves presence and SoSL, but does not exacerbate cybersickness.
- We enlargement the downward FoV, which increases the sense of feeling that the body is different from one's own body owing to tracking accuracy. This causes limited improvement in the SoA and SoBO.

7.2 Experiment 1: Line Tracing Task

7.2.1 Overview

In this experiment, we investigated whether using an HMD with an increased downward FoV can increase presence and SoE (SoSL, SoA, and SoBO) because of the improved visibility of the VE and self-avatar. The experimental setup consisted of the prototype HMD created in Chap. 6 (Fig. 7.1), a waist-mounted

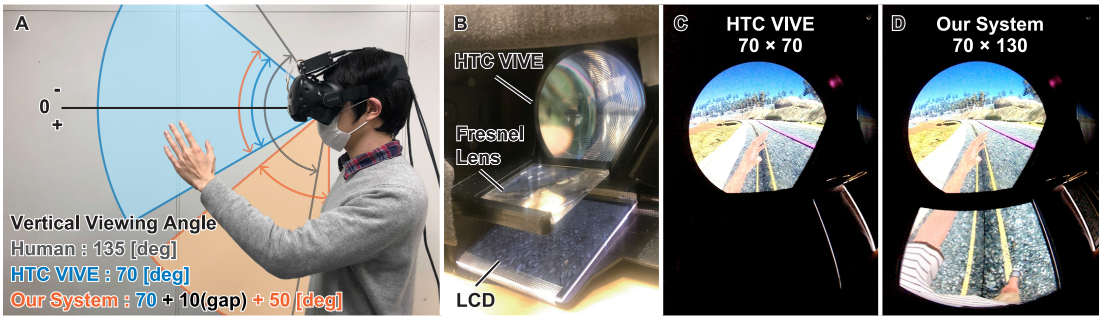


Figure 7.1: Our HMD with a wide downward FoV. A) Side view. Blue: vertical FoV of the HTC VIVE. Part of the lower FoV is missing owing to the LCDs housing. Orange: increased vertical FoV. Gray: vertical FoV of a typical human. B) The added LCD and Fresnel lens. C) Original FoV of the HTC VIVE (approx. 70×70 [deg]). The user's hand is visible but not the lower limbs and body. D) FoV of the prototype (approx. 70×130 [deg]). The user's lower limbs and body and the ground are all visible at the same time.

position tracker (HTC VIVE Tracker), and a pair of controllers (HTC VIVE Controller) with the HTC Lighthouse tracking system. In the experiment, the participants performed line tracing tasks in the VE. The VE and the course (the purple line) used in this experiment are shown in Fig. 7.2. The participants were instructed to follow the course by using the HTC VIVE controller, which goes through a dirt ground, an asphalt ground, and a shaded tunnel. The avatar moves in the VE at a running speed (approx. 11 [km/h]). Full-body avatar animations were computed by inverse kinematics (IK) using the VRIK package. Note that the participants' feet were not tracked. The avatar's feet were animated by IK from the head, hands, and waist positions. Therefore, the fidelity of the running animation is limited.

The experiment followed a 2 × 2 factorial design. The independent variables were the downward FoV (*ON*: the additional display units were used; *OFF*: the units were *not* used) and the avatar type (*Humanoid*: a full-body humanoid avatar that followed the participant's movements was displayed; *Sphere*: white spheres were displayed on the participant's hands). Thus, we have a total of four conditions (*OnH*: On_Humanoid; *OnS*: On_Sphere; *OffH*: Off_Humanoid;

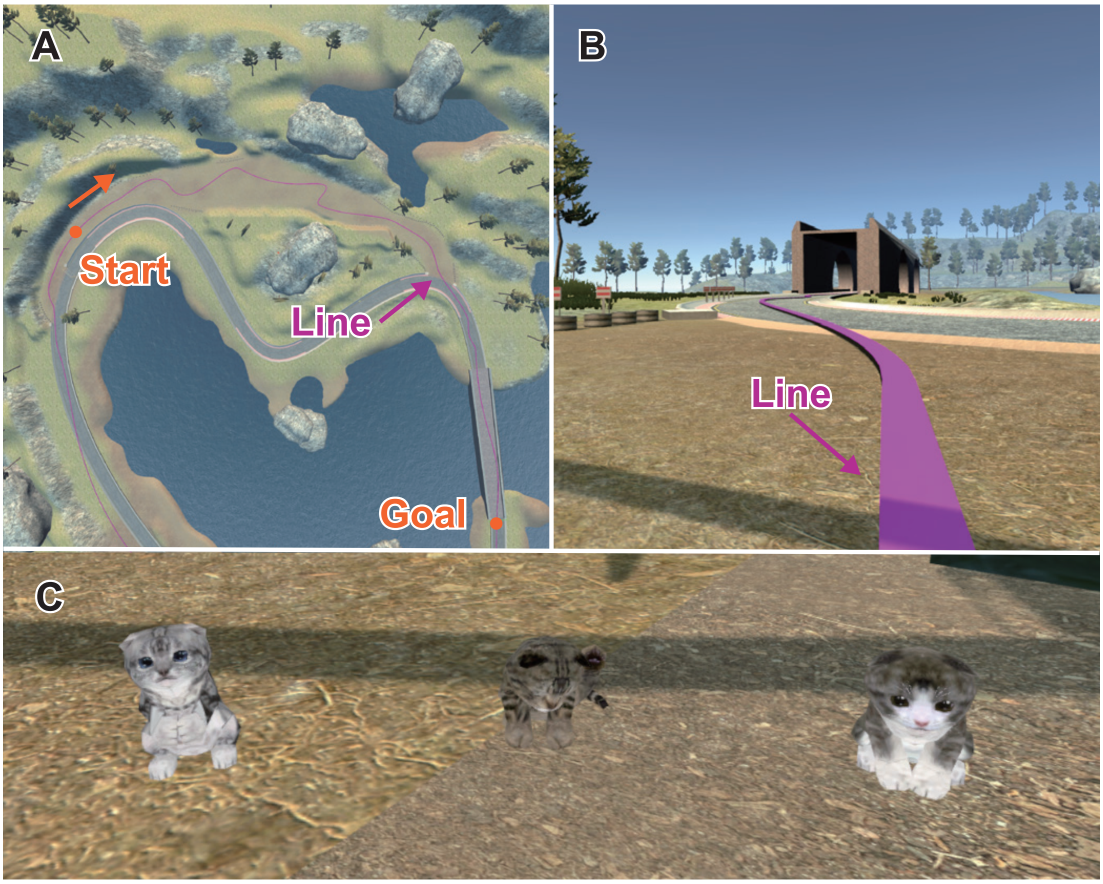


Figure 7.2: Course used in Experiment 1. A) Top view of the entire course (orange arrow indicates the direction of the course). B) Example user's view (participants were instructed to follow the purple line). C) Virtual cats to check the visibility of the downward FoV.

and *OffS*: Off_Sphere). All variables were within-subject. The participants performed the task under all the conditions in a counterbalanced order based on a Latin-square design. Figure 7.3 shows the appearance of the avatars used in the experiment. For the humanoid avatar, the participants selected a male or female Microsoft open-source avatar [143] to match their claimed gender. The sphere avatar was introduced on the basis of a previous study [99] from witch it is expected that the SoE will be lower when using a sphere avatar than when using a full-body avatar. These conditions were set for the purpose of verifying whether increasing the downward FoV increases the SoE only when a full-body avatar is

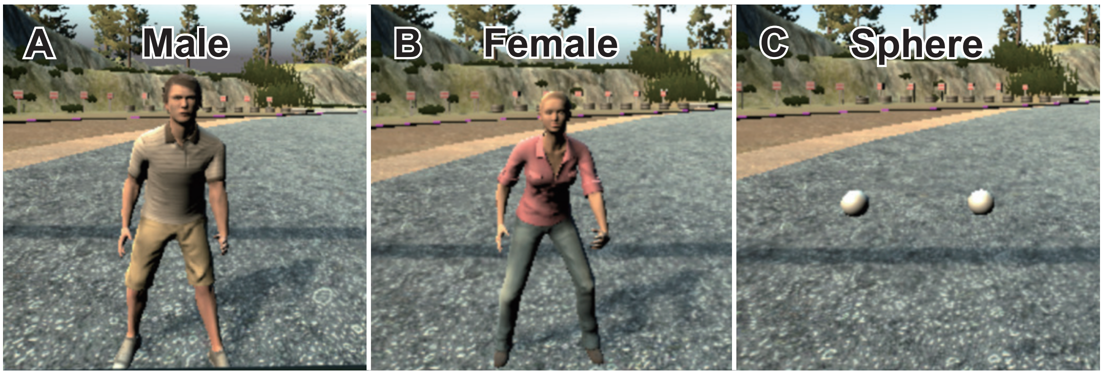


Figure 7.3: Avatars used in Experiment 1. A) Male avatar. B) Female avatar. C) Spheres displayed on the hands.

used.

7.2.2 Participants

Participants were recruited through a campus mailing list and announcements. In accordance with the ethical review committee of the author's institution, informed consent was obtained from each of the participants after the study was fully explained to them. Each participant was paid equivalent to about 10 USD. Twenty-four (sixteen males and eight females; mean age = 24.8 years; standard deviation = 3.74) participated in the study.

7.2.3 Procedure

First, the participants were shown the experimental course on a desktop monitor. They were instructed to move from the start point to the goal along the course in the VE as quickly and accurately as possible. After that, they practiced moving operations with the VIVE controller on a desktop monitor instead of the HMD. The trackpad on the left controller was used for left–right viewpoint rotation, and the trackpad on the right controller was used for forward and backward movement. Since the direction of travel followed the direction of gaze (head), the participants were able to move forward with the controller while facing the direction they wanted to go, without having to use the controller to rotate their viewpoint. The following steps were then performed under each condition.

1. Participants wear the HMD, tracker, and controllers.
2. They go in front of the mirror in the VE.
3. They move their bodies freely for 30 [s] while looking at the mirror.
4. They go near a virtual cat around the mirror (Fig. 7.2 C).
5. They interact with the cat while crouching down.
6. They move to the start point of the course.
7. They travel to the goal by following the purple line on the ground as quickly and accurately as possible.
8. They remove the HMD and complete the questionnaire.
9. They take a 3 [min] break.

Step 3 is to ensure that the participants observe the self-avatar on a virtual mirror to enhance SoA and SoBO [24, 25, 26]. Step 5 is to help them get used to looking down while wearing the HMD. Crouching motion was introduced to help them get a sense of distance from the ground. In the case of using the full-body humanoid avatar, we expected that the SoBO can be improved by being able to see their thighs and other parts of the body. The questionnaire included the IPQ [133] to investigate the presence, the Illusion of Virtual Body Ownership Questionnaire (IVBO) [144] to investigate the SoA and SoBO, the Simulator Sickness Questionnaire (SSQ) [145] to investigate the effect of increased FoV on cybersickness, and a free-text feedback. As objective measurements, we measured the elapsed time from the start to the goal (Time), the head movement pattern (Pitch Angle), and the percentage of time spent on the purple line (Score).

7.2.4 Questionnaire

The IPQ was used to evaluate presence. In this experiment, four of the subitems of the IPQ (*Pres*: General Presence; *SP*: Space Presence; *Inv*: Involvement;

and *Real*: Experienced Realism) were rated using a seven-point Likert scale [1 (strongly disagree) – 7 (strongly agree)]. *Pres, SP, Inv*, and *Real* consisted of one, five, four, and five items, respectively. In addition, in this experiment, the sum of the mean values of each evaluation item (*Total*: Total Presence) was calculated as the evaluation index.

The IVBO was used to assess SoA and SoBO. In this experiment, 10 items (*myBody, twoBodies, bodyIntensity, myMove, myMoveJoy, bodyChange, checkBody, weightBody, myExpJoy*, and *humanBody*) were measured using the seven-point Likert scale [1 (strongly disagree) – 7 (strongly agree)], excluding the question about the fire (*avoidBody*, *harmBody*, and “Why have you responded to the fire or why not?”), Free-text comments on two items (“What exactly gave you the feeling that the virtual body is your own, or what has prevented it?” and “When did the feeling of owning the virtual body was especially strong or weak?”) were collected. The meaning of each item is shown in Table 7.1. The SSQ was used for the evaluation of cybersickness. In this experiment, among the subitems of the SSQ, three (*N*: Nausea; *O*: Oculomotor; and *D*: Disorientation) were evaluated using a four-point scale [0 (None) – 2 (Severe)]. Each index consisted of seven questions. In addition, *TS* (Total Severity), which is the sum of the mean values of each item, was used as the evaluation index.

7.2.5 Hypotheses

In Experiment 1, we set the following hypotheses.

H1-1 Increasing the downward FoV improves presence, which in turn increases IPQ scores.

H1-2 Increasing the downward FoV improves the SoA and SoBO, which in turn increases the IVBO scores.

H1-3 Increasing the downward FoV makes cybersickness more likely to occur and increases the SSQ scores.

H1-4 Increasing the downward FoV causes discomfort under the On_Sphere condition owing to the lack of body display, and results in lower IPQ and IVBO than under the On_Humanoid condition.

H1-5 Increasing the downward FoV improves the SoSL, which in turn improves the percentage of time spent on the purple line (Score).

H1-6 Increasing the downward FoV improves the SoSL and shortens the task completion time (Time).

H1-7 Increasing the downward FoV improves line visibility and lowers the average downward pitch angle of the head (the head turns upward).

Table 7.1: Illusion of Virtual Body Ownership (IVBO) Questionnaire [144]. Excerpted information about the survey items.

Item	Question
myBody	I felt like the body I saw in the virtual world was my body.
twoBodies	I felt as if I had two bodies.
bodyIntensity	The illusion of owning a different body than my real one was very strong during the experience.
myMove	The movements I saw in the virtual world seemed to be my own movements.
myMoveJoy	I enjoyed controlling the virtual body I saw in the virtual world.
avoidBody	I tried to avoid touching the flames.
harmBody	In between I was worried that I might get harmed if I touched the flames.
bodyChange	At a time during the experiment I felt as if my real body changed in its shape and/or texture.
checkBody	After taking off the HMD, I felt the need to check that my body does really still look like to what I had in mind.
weightBody	I felt an after-effect as if my body had become lighter/heavier.
myExpJoy	How did you like the overall experience in the virtual world?
humanBody	I felt like the virtual body I saw looked human.

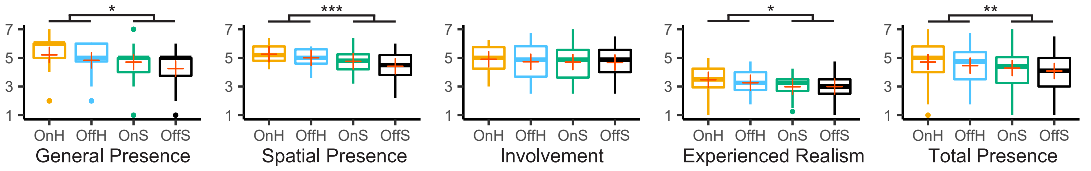


Figure 7.4: IPQ results for Experiment 1 [1 (strongly disagree) to 7 (strongly agree)].

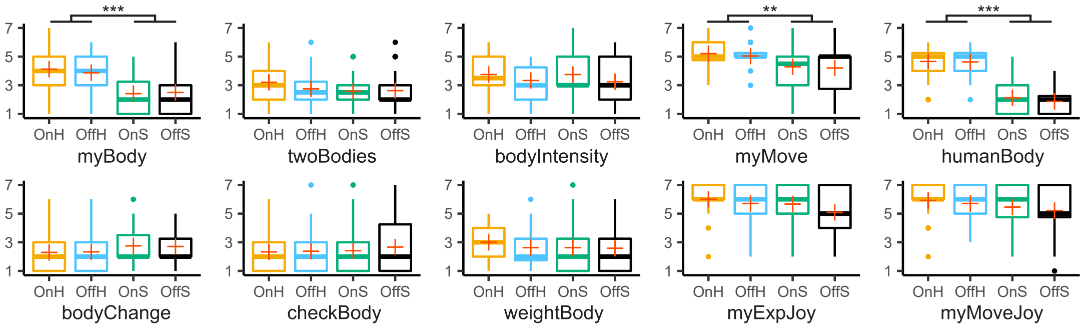


Figure 7.5: IVBO results for Experiment 1 [1 (strongly disagree) to 7 (strongly agree)].

7.2.6 Results

The results of Experiment 1 are shown in Table 7.2 and Figs. 7.4, 7.5, and 7.6. We performed two-way ANOVA (the additional LCD units: ON vs. OFF and the avatar type: Humanoid vs. Sphere) followed by a post hoc analysis with the Holm–Bonferroni correction for all results of the experiment. Significant differences are indicated with symbols (*** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$).

Table 7.2 shows the results of the IPQ, IVBO, SSQ, and Score, as well as the average time [s] from the start to the goal (Time) and the average head pitch angle [deg] (Pitch Angle). As some parts of our data (IPQ, IVBO, SSQ, Time, and Pitch Angle) were not normally distributed, we employed the aligned rank transform [127]. Two-way ANOVA did not find any significant differences in the SSQ, Time, and Pitch Angle. Figure 7.4 shows the results of the IPQ. Two-way ANOVA revealed a significant main effect of the avatar type (Humanoid vs. Sphere) on *Pres* ($p < 0.05$, $\eta_p^2 = 0.05$), *SP* ($p < 0.001$, $\eta_p^2 = 0.12$), *Real* ($p < 0.05$,

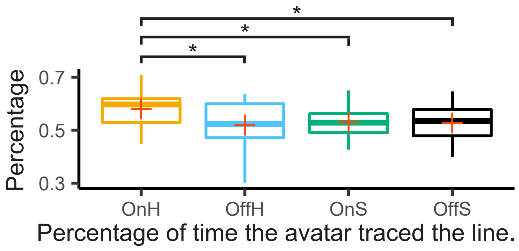


Figure 7.6: Percentage of time spent on the purple line during the task (Score) for Experiment 1.

$\eta_p^2 = 0.04$), and *Total Presence* ($p < 0.01$, $\eta_p^2 = 0.02$). No significant main effect of LCDs (ON vs. OFF) and interaction were found.

Figure 7.5 shows the results of the IVBO. Two-way ANOVA revealed a significant main effect of the avatar type (Humanoid vs. Sphere) on *myBody* ($p < 0.001$, $\eta_p^2 = 0.24$), *myMove* ($p < 0.01$, $\eta_p^2 = 0.07$) and *humanBody* ($p < 0.001$, $\eta_p^2 = 0.58$). No significant main effect of LCDs (ON vs. OFF) and interaction were found.

Figure 7.6 shows the results of the percentage of time spent on the purple line (Score). Two-way ANOVA revealed a significant main effect of LCDs (ON vs. OFF) on the Score ($p < 0.05$, $\eta_p^2 = 0.16$). In addition, we found an interaction effect ($p < 0.05$, $\eta_p^2 = 0.19$). Post-hoc analysis using a Holm method found significant differences between *OnH* and *OffH* ($p < 0.05$, $d = 0.80$), *OnH* and *OnS* ($p < 0.05$, $d = 0.82$), and *OnH* and *OffS* ($p < 0.05$, $d = 0.80$).

7.2.7 Discussion

The main findings are as follows.

1) In terms of presence, the participants perceived higher presence when using the full-body humanoid avatar than when using the sphere avatar (*Pres*, $p < 0.05$; *Sp*, $p < 0.001$; *Real*, $p < 0.05$; and *Total Presence*, $p < 0.01$). There was no significant difference in perceived presence between the LCD On and Off conditions. Thus, [H1-1] and [H1-4] were not supported.

2) As for SoA and SoBO, the participants perceived higher SoA and SoBO when using the full-body humanoid avatar than when using the Sphere avatar (*myBody*, $p < 0.001$, *myMove*; $p < 0.01$, and *humanBody*; $p < 0.001$). However, there was

no significant difference between the LCD On and Off conditions, which did not support [H1-2]. Under the condition where the downward FoV was increased and the humanoid avatar was used (*OnH*), the participants gave positive opinions regarding the visibility and connectivity of the body, such as “I felt a sense that

Table 7.2: Experimental results in Experiment 1. Each value represents the mean and the standard deviation. Items with a significant difference are in bold.

		OnH	OffH	OnS	OffS
IPQ	**Pres**	**5.21±1.25**	**4.83±1.17**	**4.71±1.37**	**4.25±1.48**
	Sp	**5.26±0.65**	**5.01±0.61**	**4.76±0.80**	**4.41±0.94**
	Inv	4.91±0.90	4.73±1.29	4.71±1.26	4.68±1.21
	Real	**3.47±1.03**	**3.26±0.89**	**2.98±0.79**	**2.99±0.95**
	Total Presence	**4.64±0.55**	**4.42±0.61**	**4.23±0.68**	**4.07±0.82**
IVBO	**myBody**	**4.13±1.51**	**3.88±1.19**	**2.42±1.32**	**2.50±1.50**
	twoBodies	3.21±1.38	2.75±1.36	2.58±1.25	2.63±1.44
	bodyIntensity	3.75±1.45	3.33±1.24	3.75±1.51	3.25±1.67
	myMove	**5.21±1.10**	**5.04±0.91**	**4.29±1.71**	**4.21±1.79**
	myMoveJoy	5.92±1.21	5.71±1.27	5.46±1.50	5.21±1.77
	bodyChange	2.29±1.30	2.33±1.37	2.75±1.57	2.71±1.33
	checkBody	2.33±1.49	2.38±1.66	2.42±1.72	2.67±1.93
	weightBody	3.00±1.35	2.63±1.50	2.63±1.84	2.58±1.61
	myExpJoy	6.00±1.14	5.71±1.37	5.67±1.43	5.08±1.79
	humanBody	**4.67±1.27**	**4.63±1.13**	**2.13±1.08**	**1.88±0.99**
SSQ	Nausea	1.58±2.96	1.75±2.85	1.54±2.08	1.29±2.12
	Oculomotor	3.88±4.74	3.83±4.16	4.08±4.17	3.21±3.99
	Disorientation	2.63±3.52	2.33±3.06	2.67±3.80	2.21±3.15
	Total Severity	5.96±7.80	5.92±6.95	6.29±7.14	5.04±6.53
Score	**Score**	**0.58±0.06**	**0.52±0.09**	**0.53±0.06**	**0.53±0.07**
Time	Time	177.34±4.38	176.88±7.19	176.32±7.51	175.66±4.76
Head	Pitch Angle	0.80±8.17	2.04±5.88	-0.77±5.78	-0.18±6.99

the arms and legs were connected to my body." and "I felt that it was my body when I could see clearly my hands and below." However, there were also some negative comments, such as "the position of the shoulder was unnatural" and "I felt uncomfortable that my legs did not move in accordance with my movements." It is possible that the increase in the downward FoV made it easier to notice the inaccuracy of body tracking, which inhibited the SoA and SoBO. In addition, for most of the experiment, the participants performed the task of controlling the avatar with the controller without moving their bodies significantly, which may have prevented them from taking advantage of the increased downward FoV. There was no significant difference in perceived SoA and SoBO between the LCD On and Off conditions when the participants used the Sphere avatar. Thus, [H1-4] was not supported.

3) There was no significant difference in cybersickness among conditions. Therefore, the increased downward FoV did not increase cybersickness in this experiment, and the results did not support [H1-3]. In a previous study that showed the relationship between the increase in horizontal FoV and cybersickness, a 2 [min] task was used [87]. In this experiment, the time required to execute the task was about 3 [min] which is considered sufficiently long to verify the effect on cybersickness. In a previous study [87], Lin *et al.* confirmed that there was a large difference in the SSQ Score between small (60 [deg], 100 [deg]) and medium (140 [deg]) horizontal FoVs. They also confirmed that the SSQ Score does not change significantly between the medium (140 [deg]) and the large (180 [deg]) FoVs. These results suggest that the effect of an increased FoV of the HMD on cybersickness is marginal when it is wider than a certain range.

4) For the self-positioning ability, the participants tended to trace the line more accurately with the increased downward FoV than with the normal FoV (Score, $p < 0.05$). An interaction between LCD and Avatar was found, $p < 0.05$, and the participants tended to trace more accurately when using the full-body humanoid avatar with the increased downward FoV (*OnH*) than under the other conditions (between *OnH* and *OffH*, $p < 0.05$; between *OnH* and *OnS*, $p < 0.05$; and between *OnH* and *OffS*, $p < 0.05$). Dewez *et al.* reported that the presence or absence of an avatar did not change the performance of the line tracing task [112]. However, the results of Experiment 1 showed that the combination of a full-body

humanoid avatar and the increased downward FoV improved the performance of the line tracing task. Unlike the sphere avatar, the humanoid avatar's body in the downward FoV was in contact with the ground, which seems to have made it easier for the participants to recognize where they were standing. This combination also improved the SoSL, supporting [H1-5]. On the other hand, there was no significant difference in the task completion time (Time) between the conditions. Thus, [H1-6] was not supported despite the finding that the perceived SoSL was improved. Table 7.2 shows that no significant difference was found in the average pitch angle of the head during the experiment. This did not directly support [H1-7]. When comparing the LCD On and Off conditions, there was a tendency for the head to be tilted more slightly upward (smaller angles) when with the increased downward FoV. Some participants commented that they no longer needed to concentrate on their feet when with the increased downward FoV. Behavioral observations and participants' comments after the experiment suggest that some participants under the *OnH* condition checked their bodies more frequently than necessary owing to the novelty of seeing the self-avatar in the lower FoV. This could have made it more difficult to measure differences in head movement.

Our results confirmed that the use of the full-body humanoid avatar improved presence, SoA, and SoBO compared with the use of the sphere avatar. This finding differs from those of the study by Lugrin *et al.* where they compared controller and upper body avatars [104], suggesting that the increased downward FoV may have improved the visibility of the self-avatar and made the difference between avatar conditions more pronounced. On the other hand, under the condition where the same avatar was manipulated, no increases in the presence, SoA, and SoBO were observed owing to the increase in the downward FoV. This may have had a negative effect on the presence, SoA, and SoBO owing to the fact that the increased FoV made the low tracking accuracy more pronounced and made it easier to see the self-avatar that looked different from oneself. On the other hand, it is also possible that no significant difference was found owing to the fact that there was little time for the avatar to appear in the lower visual field owing to the lack of body movement during the experiment. Therefore, we conducted Experiment 2, which included a task that required large body and hand movements, to examine the effect of increasing the amount of time the avatar was visible in the downward FoV.

7.3 Experiment 2: Escaping Task

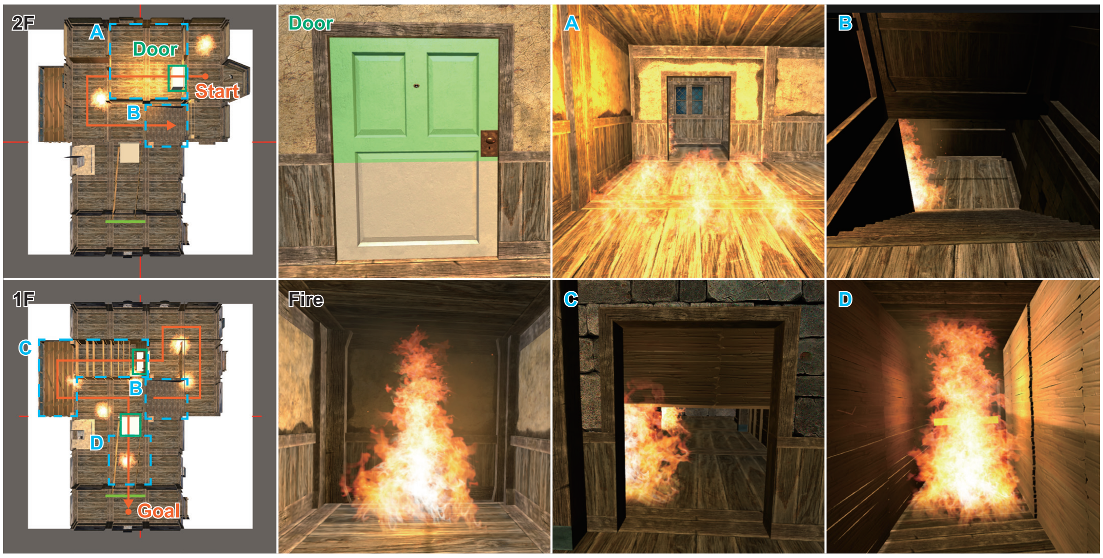


Figure 7.7: Course used in Experiment 2. Orange lines represent the course from the start to the goal. 2F) The second floor of the course. 1F) The first floor of the course. Door) A door that opens when touched by the avatar's hand for about 3 [s]. Fire) Avoidable tall fire on the course at about the head height. A) Avoidable short fire at about the shin height. B) Descending stairs from 2F to 1F. C) Obstacle near the head. D) Unavoidable tall fire at about the head height.

7.3.1 Overview

In Experiment 2, we set up a task requiring large physical movements (using hands to open doors, avoiding tall fires, descending stairs, squatting, and touching large fires) and investigated the effect of increasing the downward FoV on presence, SoA, and SoBO with the full-body humanoid self-avatar. The experimental course created in Unity is shown in Fig. 7.7. The equipment and the humanoid avatars are the same as those in Experiment 1. In Experiment 1, there was no interaction between the LCD conditions and the avatar conditions in terms of presence, SoA, and SoBO ratings. Therefore, in Experiment 2, we excluded the Sphere avatar conditions in Experiment 1 and left only two experimental conditions (*OnH*: On_Humanoid; and *OffH*: Off_Humanoid) using the full-body humanoid avatars. Participants in the experiment performed the task under experienced all the conditions in a counterbalanced order.

The course consisted of a two-story building, and participants were instructed to use a controller to move from the start point on the second floor to the goal on the first floor. A door was placed along the course, which disappeared when the participant touched it with their hand for about 3 [s], allowing him/her to move on (Fig. 7.7, Door). This action requires the participant to thrust his or her hand forward in order to touch the door naturally, which is introduced to implicitly confirm the connection between the hand and the upper arm. Fire objects were placed at several locations along the course (Fig. 7.7, Fire). It is known that the higher the SoBO, the higher the tendency for users to avoid dangerous objects such as fire [99, 144, 146], and this tendency may become more pronounced when the downward FoV is increased. A short fire was placed at the beginning of the course (Fig. 7.7 A). The purpose of this fire was to test whether patterns of head movements of the participants would change when they could see obstacles under their feet by increasing the downward FoV. The participants had to descend a staircase (Fig. 7.7 B) to reach the goal. This was added to determine whether the increase in the downward FoV affects the head angle when descending the stairs. In addition, wooden obstacles were placed at the height of the head at some points along the course (Fig. 7.7 C). The participants could avoid the obstacles by crouching down, but they could also pass through them in an upright position, simply because they were unable to see ahead. This was set up for the purpose of investigating whether the participants' patterns of head movements would change with the change in the distance between the ground and the head by crouching and the change in the visibility of the feet when crouching. In addition, an unavoidable fire was placed near the goal (Fig. 7.7 D) to determine whether the number of effects displayed from the ground increases as the downward FoV increases, thereby increasing the sense of realism.

7.3.2 Procedure

The participants were the same as those in Experiment 1. They participated in Experiment 2 on a later day than Experiment 1. Each participant was paid equivalent to about 10 USD. After the participants were briefed about the experiment and signed the experimental consent form, they checked the course on a desktop monitor. At this time, they were instructed that they were required to move from the start point to the goal of the course while wearing the prototype HMD. After that, the participants wore the HMD, tracker, and controllers and performed a practice travel from the start to the goal using the VIVE controller (the same control method as in Experiment 1) under the Off_Sphere condition of Experiment 1. The purpose of the practice travel was to have the participants roughly memorize the route, so no fire was set up along the course. The speed of the travel was adjusted to a walking speed (approx. 4 [km/h]). After arriving at the goal, the HMD was removed, and after a 3 [min] break, the following steps were performed under each condition.

1. Participants wear the HMD, tracker, and controllers.
2. They go in front of the mirror in the VE.
3. They move their bodies freely for 30 [s] while looking at the mirror.
4. They freely observe their own body and the surrounding objects (virtual cats, fire) for 30 [s].
5. They move to the start point of the course.
6. They travel to the goal.
7. They remove the HMD and complete the questionnaire.
8. They take a 3 [min] break.

In Step 4, participants encountered the same virtual cat as in Experiment 1 and a fire at the height of their head (Fig. 7.7, Fire). At this time, the participants were told to “freely observe your own body and surrounding objects for 30 [s]” and were not required to touch the fire. The questionnaire consisted of the same items as those in Experiment 1 (IPQ, IVBO, and SSQ) with the addition of questions related to the fire in the IVBO (*avoidBody, harmBody*, and “Why have you responded to the fire or why not ?”). We measured the elapsed time from the start to the goal (Time), the position in the VE, and the pitch angle of the head as the objective measurements.

7.3.3 Hypotheses

In Experiment 2, we set the following hypotheses on the basis of the results of Experiment 1.

H2-1 Increasing the downward FoV improves the IPQ scores.

H2-2 Increasing the downward FoV improves the IVBO scores.

H2-3 Increasing the downward FoV increases the SSQ scores.

H2-4 Increasing the downward FoV enhances the fire-avoidance behavior by increasing the SoBO, and increases the elapsed time from the start to the goal (Time).

H2-5 Increasing the downward FoV reduces the downward pitch angle of the head by improving the visibility of obstacles at the feet (Fig. 7.7 A).

H2-6 Increasing the downward FoV reduces the downward pitch angle of the head, specifically when descending the stairs (Fig. 7.7 B).

7.3.4 Results

The results of Experiment 2 are shown in Table 7.3 and Figs. 7.8 and 7.9. We performed the Wilcoxon signed-rank test (LCD conditions: ON; *OnH* vs. Off; *OffH*). Significant differences are indicated with symbols (*** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$).

Table 7.3 shows the results of the IPQ, IVBO, and SSQ, as well as the average time from start to goal (Time) [s] and the average pitch angle of the head at the locations corresponding to those in Fig. 7.7 A, B, C, and D [deg]. The Wilcoxon

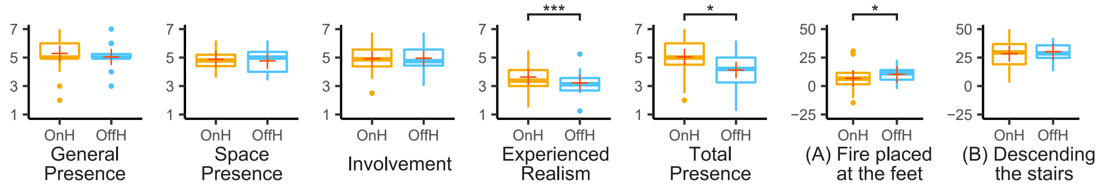


Figure 7.8: IPQ results and average head pitch angle at each location in Experiment 2. The five graphs from the left show IPQ results [1 (strongly disagree) to 7 (strongly agree)]. The two graphs from the right show the average head pitch angle at each location. Only important items corresponding to Fig. 7.7 A and B are shown (0 [deg] when facing front and +90 [deg] when fully looking down).

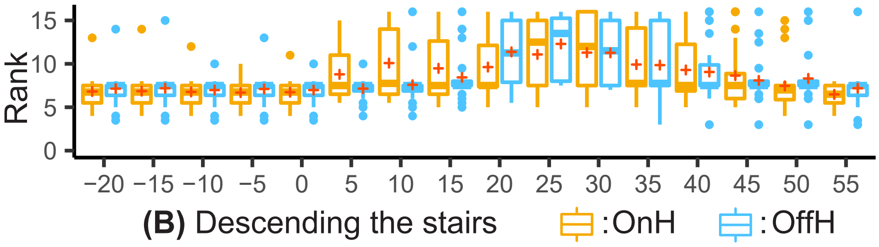


Figure 7.9: Histogram of the frequency of occurrence of the head pitch angles at the location corresponding to Fig. 7.7 B in Experiment 2, ranked into 16 levels in 5 [deg] intervals (Orange: *OnH*; Blue: *OffH*). The vertical axis indicates the rank (the higher the more frequent) of each interval. The horizontal axis indicates the head pitch angle intervals (($X \leqq X_{raw} < X + 5$) at X [deg]). The head pitch angle is 0 [deg] when facing front and +90 [deg] when fully facing down.

signed-rank test did not show any significant differences in IVBO, SSQ, or Time. A significant difference in IVBO (*myMove*) may be found with a larger number of participants ($p = 0.053$, $r = 0.41$). Figure 7.8 (left) shows the results of the IPQ. The Wilcoxon signed-rank test showed significant differences in *Real* ($p < 0.001$, $r = 0.64$) and *Total Presence* ($p < 0.05$, $r = 0.46$). Figure 7.8 A and B shows the results of the average pitch angle of the head at the locations corresponding to those in Fig. 7.7 A and B. The vertical axis indicates the pitch angle of the head, which is 0 [deg] when facing forward and +90 [deg] when facing down completely. The Wilcoxon signed-rank test showed a significant difference in Fire placed at the location (A) ($p < 0.05$, $r = 0.44$). Figure 7.9 is a histogram of the frequency of occurrence of each 5 [deg] interval of the head pitch angle at the location corresponding to that in Fig. 7.7 B and ranked in 16 levels. 0 [deg] on the horizontal axis indicates the range of ($0 \leqq x < 5$) [deg].

Table 7.3: Experimental results in Experiment 2. Each value represents the mean and standard deviation. Items with a significant difference are in bold.

		OnH	OffH
IPQ	Pres	5.29±1.23	5.04±0.91
	Sp	4.88±0.70	4.77±0.86
	Inv	4.92±1.03	4.96±0.96
	Real	**3.64±0.95**	**3.22±0.83**
	Total Presence	**4.56±0.45**	**4.40±0.47**
IVBO	myBody	3.79±1.41	3.83±1.27
	twoBodies	3.21±1.38	3.50±1.47
	bodyIntensity	3.79±1.44	3.79±1.28
	avoidBody	4.04±2.39	4.13±2.33
	harmBody	3.71±2.16	3.63±1.88
	myMove	5.04±0.91	4.50±1.18
	myMoveJoy	6.04±0.75	5.92±0.97
	bodyChange	2.58±1.38	2.71±1.33
	checkBody	2.42±1.50	2.25±1.42
	weightBody	2.42±1.47	2.58±1.72
	myExpJoy	6.04±0.86	5.92±1.06
	humanBody	4.29±1.20	4.08±1.28
SSQ	Nausea	0.96±1.83	0.88±1.48
	Oculomotor	1.33±2.14	1.63±2.26
	Disorientation	1.00±1.47	1.08±1.47
	Total Severity	2.50±3.78	2.83±3.52
Time	Time	77.43±16.27	77.34±11.85
Head	**Pitch Angle (A)**	**6.77±10.07**	**10.25±6.49**
	Pitch Angle (B)	28.38±12.47	30.20±7.85
	Pitch Angle (C)	4.98±9.59	4.85±7.88
	Pitch Angle (D)	4.06±10.08	5.69±7.06

7.3.5 Discussion

The main findings are as follows.

1) Unlike in Experiment 1, the participants perceived higher presence with the increased downward FoV than with the normal FoV (*Real*, $p < 0.001$; and *Total Presence*, $p < 0.05$). In particular, a large difference was detected in the item "measuring the subjective experience of realism in the VE," suggesting that the FoV becomes closer to the real environment by increasing the downward FoV (*Real*, $p < 0.001$). Compared with Experiment 1, we speculate that self-image was more easily perceived owing to the larger self-motion; thus, presence was increased in Experiment 2. These results supported [H2-1].

2) For the SoBO, no significant differences were detected for all items, despite the increase in body movements in Experiment 2, and the results do not support [H2-2]. Some participants commented that they felt uncomfortable with the elbow and head movements (*OnH*), suggesting the need for elbow and shoulder tracking when the downward FoV is increased. Another participant said, "I felt that my body was burning more in *OnH* than in *OffH* on the fire, but I strongly felt that I was in the VE because there was no heat feedback." This comment suggests that multimodal feedback that is coherent with the content becomes more important when the downward FoV is increased. The item "the movements I saw in the virtual world seemed to be my own movements," corresponding to the SoA, may show a significant difference with a larger number of participants (*myMove*, $p = 0.05321$, $r = 0.41$), which encourages further investigation. On the other hand, one participant commented, "the amount of information coming from the front monitor was so large that I did not need to pay attention to the lower monitor (in *OnH*)," suggesting the limitation of information presented to the peripheral vision.

3) Regarding cybersickness, no significant difference was found between *OnH* and *OffH* as in Experiment 1. Thus, [H2-3] is rejected. Even though we did not analyze cybersickness or its sign may be detected after the experiment is over [147] or by observing head posture instability [148].

4) The tendency to avoid the fire did not change with the increase in the downward FoV; thus, [H2-4] was not supported. The participants' comments suggest that there were roughly two types of behavior with the increased downward FoV.

Some participants “felt fear of fire” and “noticed the fire and avoided it to a greater extent than when under *OffH*,” whereas others “noticed that fire was not scary” and “got closer to the fire and did not have to avoid it as much as when under *OffH*.” The former group’s behavior seems in line with that observed in previous studies, which showed the tendency to avoid obstacles farther with increasing realism [107, 116]. It is possible that the increase in the downward FoV increased the overall perceived realism and made the obstacles seem more like real objects. On the other hand, the latter group’s behavior suggests that inconsistent feedback (e.g, lack of heat) was more apparent, thus the perceived realism was reduced. As such, we found significant individual differences in the tendency to avoid obstacles with the increase in the downward FoV, which need to be investigated in more detail in the future.

5) The mean pitch angle of the head in the presence of a fire at the feet (Fig. 7.8 A) tended to be lower (i.e., more upward) when the downward FoV was increased (*OnH*, $p < 0.05$), supporting [H2-5]. This suggests that the increase in the downward FoV makes it possible to see obstacles under the feet without having to turn the head downward too much. On the other hand, Fig. 7.8 B shows that no significant difference was detected in the mean pitch angle of the head when descending the stairs. However, Fig. 7.9 shows that the *OffH* histogram is unimodal with 25 [deg] at the top, whereas the *OnH* histogram is bimodal with 10 and 30 [deg] at the top. This indicates that some participants did not markedly turn their heads downward when descending the stairs with the increased downward FoV. This result partially supports [H2-6].

In Experiment 2, we speculate that the participants had to turn their heads downward more without the downward FoV to see the hazards (fire and descanting stairs) that they could not see otherwise. However, in the case of the stair, there was no significant difference between the two LCD conditions. We speculate that this is because they had to see the downstairs to move down even with the downward FoV.

We also found that presense tended to improve without any adverse effect on cybersickness when the downward FoV was increased. In other words, the increase in the downward FoV is likely to have a positive effect on the VR experience. However, it had little impact on the SoBO and SoA. This may be due to the fact that an increased downward FoV does not only have a positive effect, but it also reveals negative information such as low tracking accuracy, the self-avatars that look different from their own, and lack of multimodal feedback. In other words, more advantages may be found by using accurate body tracking, a humanoid avatar that looks like the user, and a coherent multimodal feedback. We also observed that the participants tended to be able to see the obstacles and the descending stairs without having to significantly turn their heads downward. This suggests that the head movements were closer to that in the real environment, which may have also contributed to the improved presence.

7.4 Overall Discussion

Results of Experiment 2 confirmed that increasing the downward FoV improved the presence of the VR experience. In addition, the head pitch angle tended to be upward when the downward FoV was required such as descending the stairs. More natural head movement may make the overall experience more realistic. Situations that will benefit from the increased downward FoV include holding a baby in the arms, an animal approaching the feet, and walking in a high place. In Experiment 1, increasing the downward FoV improved the SoSL and the performance of the line tracing task when using a full-body humanoid avatar. Unlike studies using a full-body humanoid avatar with a limited downward FoV [112], the combination of an increased downward FoV and the display of a full-body humanoid avatar made it easier to recognize where one was currently standing. In addition, there was no adverse effect on cybersickness, which was a concern with the wider FoV. However, the task duration of the experiment in this study was about 3 [min], and it is necessary to investigate the effects of prolonged use.

In Experiment 1, we observed a tendency for the condition with the full-body humanoid avatar to improve presence, SoA, and SoBO compared with the condition with the spherical avatar. Lugrin *et al.*'s study [104] comparing controller and upper body avatars did not show these changes, but it is possible that the difference between the avatar conditions became more pronounced as a result of the increased visibility of the avatar owing to the increased downward FoV. On the other hand, when we compared the effects of using a full-body humanoid avatar with or without a lower visual field in both Experiments 1 and 2, the change in the SoA and SoBO was limited. There were some participants who felt uncomfortable owing to the discrepancy between the avatar appearance and their proprioceptive sensation as a result of the increased visibility of the avatar's shoulders and elbows. We also believe that the discrepancy between the displayed virtual body and the actual body may have inhibited the increase in SoE and exacerbated the cybersickness. In this study, we used inverse kinematics to compute full-body animations by tracking the head, hands, and waist. However, the tracking locations and accuracy of the shoulders and upper arms may have been inadequate; therefore, more accurate whole-body tracking may increase the SoA and SoBO.

On the other hand, even if highly accurate full-body tracking is performed and the downward FoV is enlarged, the SoA and SoBO may not be improved owing to the limitations of human downward FoV and resolution. For example, even in the real environment where there is no restriction on the FoV, the downward FoV of human vision is approx. 70–80 [deg], and only a small amount of the body can be seen when looking straight ahead. In addition, since the ability to receive information in the peripheral visual field is lower than that in the central visual field [92], such information in the periphery may not be noticed.

One potential application of the HMD with an increased downward FoV is the Proteus effect [32], which refers to the influence of the avatar on the user's mental and behavioral status. Unlike previous studies, there is a possibility that the Proteus effect is induced more stably without a virtual mirror with the increased downward FoV since the self-avatar is always visible to some extent. On the other hand, being able to see the self-avatar all the time may not be helpful for the Proteus effect as it is in the periphery, or even harmful because of imperfect tracking and registration. Further investigation is necessary on the effect of the increased downward FoV on the Proteus effect.

In the present study, IVBO was used to measure the effect on the SoE, but these measurements have the problem of being easily influenced by subjectivity. Currently, a study [149] is underway to create a questionnaire with higher power to investigate embodiment in detail, and the use of a new index may allow an in-depth investigation of the effect of increasing the downward FoV on the SoE.

Semi-structured interviews can be used to investigate more detailed effects on bodily sensations.

In Experiment 2, some participants felt uncomfortable because they could not feel the heat as a result of the improved visibility of the fire. It is possible that the lack of other sensory information, such as heat and smell, which are more clearly perceived when you are closer to the object in reality, is more noticeable because nearly objects are more visible with the improved downward FoV.

One limitation of our prototype is that there was a non-negligible gap between the front and bottom displays of around 10 [deg] owing to the 3D printed housing, which prevented a unified borderless FoV. Although none of the participants reported that this gap had a negative impact, and it could even be regarded as glasses, reducing the gap could improve the overall experience, which may also yield an improved SoBO. We will develop an improved system using dedicated eyepieces such as a fused lens [‖]. As another hardware issue, the eyebox may be narrower than a normal HMD owing to the use of multiple lenses and LCDs, although none of the participants reported that the lower FoV became invisible as the eyes rotated, etc. Owing to the limited space of the HMD, the lenses and LCDs are fixed in the current prototype, but we plan to add an interpupil distance adjustment mechanism. We have also adjusted the intrinsic and extrinsic parameters of the virtual camera manually, but the lower FoV appeared slightly distorted for some participants. Even though only a few of the participants reported the distortion throughout the experiments, it should be further minimized and its impact should be investigated because it worsens the participant's perception as well as the body tracking problem. We also plan to add an eye tracker in the future to investigate how much the participants actually use the downward FoV.

7.5 Conclusion

In this chapter, we developed a VR-HMD with an increased downward FoV and examined its effects on presence, SoE (SoSL, SoA, and SoBO), cybersickness, and

[‖] Panasonic, `https://channel.panasonic.com/contents/19737/`, last accessed March 17, 2023.

head movement patterns in items of being able to see the feet, upper body, and upper arms of the self-avatar, which are difficult to present with a conventional HMD. The results of the experiments showed that increasing the downward FoV improved the presence and SoSL, that it did not have any adverse effect on cybersickness, and that it was possible to see the obstacles at the feet without turning the head downward. On the other hand, the effects on the SoBO and SoA were limited. It was suggested that the gap between the proprioceptive sensation and the avatar appearance in the VE (especially shoulders and elbows) tended to be perceived as uncomfortable and that the difference between the appearance of the real body and the avatar and the lack of heat sensation when approaching a fire could reduce presence and the SoE. In these experiments, because of the limited space, we displayed a full-body avatar with full-body animation using inverse kinematics by tracking only the head, both hands, and waist. High-precision tracking equipment can improve the sensation, including the SoE.

In the future, we will conduct follow-up studies on the effects of an improved HMD with a negligible gap between display units, precise whole-body tracking, and coherent multimodal feedback.

These results indicate that the HMD developed in Chap. 6 with an increased downward FoV can display avatars and change their appearance. In this doctoral thesis, it was not possible to investigate the influence of the avatar display on multisensory flavor perception due to time constraints. We plan to conduct additional experiments in the future using HMDs with an increased downward FoV to improve the likelihood of visual information by displaying avatars and to change multisensory flavor perception by changing the avatar's appearance.

8 Conclusion

8.1 Findings

In this dissertation, we proposed a method to visually change the appearance of the food and the surrounding environment to change the multisensory flavor perception of the food being eaten. Compared with conventional systems, our proposed visual modulation systems can modulate visual images for various food types and complex deformations of food by using machine learning techniques.

In Chap. 3, we developed a gustaory manipulation system that uses StarGAN to change the appearance of food into the appearance of a different food, successfully changing the taste and type of food to the intended food. In Chap. 4, we found that the change in perception of the taste and type of food by the gustatory manipulation system that changes the appearance of the food developed in Chap. 3 was persistent in many participants. Their persistent gustatory changes are divided into three groups: those in which the intensity of the flavor perception change gradually increased, those in which it gradually decreased, and those in which it did not fluctuate, each with a similar number of participants. These results also revealed that those participants who are less familiar with the original and target types of food feel stronger gustatory manipulation.

In Chap. 5, we developed and investigated the effectiveness of Ukemochi which detects and tracks the food segmentation image and overlays on the VEs. We found that displaying only segmented food images overlaid on the VE gives the user a higher presence compared to displaying original video frames or no food images in the VE. We also found that the food segmentation images had a similar level of ease of eating compared to the original images. On the other hand, the taste and appearance of the food were not affected favorably perhaps due to the problems of low resolution and detection accuracy.

In Chap. 6, an HMD with an increased downward FoV was developed to solve the problem of not being able to present images near the mouth. we investigated whether the increased visibility of food near the mouth due to an enlarged downward FoV improves the ease of eating and the amount of the cross-modal effect by visual modulation. Experimental results show that HMDs with an increased downward FoV can still manipulate taste of food, as in the case of Chap. 3. However, we did not find significant differences in the amount of cross-modal effects and ease of eating in the experiment with increasing downward FoV. Whereas, we found that changes in the appearance of food, with or without expansion of the increased downward FoV, can change the perception of smell and food texture to the intended food.

In Chap. 7, we investigated the effects of displaying the avatar's upper body and upper arm in the downward FoV as a preliminary step before investigating the effects of the avatar's visual presentation and appearance on multisensory flavor perception. Previous related research examining the effectiveness of avatar displays used existing HMDs with limited downward FoV, making it difficult to see the avatar's body except for its hands unless a virtual mirror installed in the VE is used. We experimented with displaying avatars in the downward FoV using HMDs with an increased downward FoV angle created in Chap. 6. As a result, it was clarified that the HMD with an increased FoV improved presence and SoSL. Also, it was confirmed that the user could see the object below with a head movement pattern close to the real behavior, and did not suffer from cybersickness. Moreover, the effect of the increased downward FoV on SoBO and SoA was limited since it was easier to perceive the misalignment between the real and virtual bodies. We have successfully developed the visual modulation system necessary to investigate the effects of avatar display and appearance changes on multisensory flavor perception.

8.2 Limitations and Future Work

8.2.1 Taste the foods we want to eat

In this dissertation, it was shown that by changing the appearance of food to that of a different food, it is possible to perceive the taste, type of food perception, smell, and food texture of the intended food. These results indicate that visual modulation can manipulate the type of food currently being eaten. Users can change the food currently being eaten by manipulating the food type. For example, if users “want to eat a food but cannot” because of diet, illness, or religious reasons, they may still be able to eat the food they want.

On the other hand, the amount of multisensory flavor perception changes by our gustatory manipulation system is insignificant, and the effect is limited to a slight change in flavor or aftertaste We believe adding an olfactory display is the most effective way to increase this effect size. When we perform multisensory integration, the likelihood of the information presented is important [29, 36]. Since the sense that has the most significant influence on multisensory flavor perception is smell [37], we believe that the presentation of olfactory information improves the likelihood and the amount of effect on gustation. Whereas, we expect that the gustatory manipulation by the presentation of olfactory information has a minor effect on the manipulation of food type (i.e., what is being eaten). For example, matcha-flavored foods that can be easily purchased at supermarkets taste like matcha, but the effect of feeling that one is drinking matcha is small. As shown in the results of Chap. 3 and Chap. 4, the effect of our gustatory manipulation system on the perception of food type is more significant than the change in multisensory flavor perception. Therefore, the weight of the senses on the likelihood of recognizing the food eaten type is more significant for the visual than for the olfactory. We believe that the combination of visual modulation and olfactory presentation creates a synergistic effect, allowing users to modify further the food they are currently eating at will.

8.2.2 Multimodal stimulus

The problem with combining current olfactory displays with HMDs is that many are mounted at the bottom of the HMD, covering the user's nose and mouth and thus inhibiting eating behavior. The HMD with an increased downward FoV developed in Chap. 6 that enables the presentation of visual information near the mouth has the same problem. In order to achieve both visual modulation and olfactory presentation, it is necessary to consider the presentation method of both. Whereas the experimental results in Chap. 6 revealed that visual modulation could change the smell of food. We believe that improving the quality of visual modulation may be possible to manipulate gustatory without relying on olfactory presentation.

Most studies that use cross-modal effects to change food texture have involved changing auditory. Zampini *et al.* altered food texture by presenting auditory information that changed the chewing sound of eating wet potato chips to a crunchy sound [42]. Koizumi *et al.* applied this effect to make daifuku (sweet red bean rice cake) feel sticky [5]. In these studies, the chewing sound measured by a microphone attached to the mouth is presented to the user by changing its frequency using a high-pass filter, thereby changing food texture. Whereas, the experimental results in Chap. 6, which changed the appearance of somen noodles into fried noodles, revealed that visual modulation changed the food texture of the noodles. Comparing somen noodles and fried noodles, the food used in the experiment, fried noodles are more crunchy because they contain brine. Participants may have perceived this difference in crunchiness as a change in texture. In the experiment in Chap. 3, in which the appearance of the somen noodles was changed to ramen noodles, participants who felt as if straight noodles changed to frizzy noodles were confirmed. The effect of visually changing the appearance of the food has the potential to produce a more flexible and delicate food texture than the conventional method of changing the frequency of chewing sounds.

8.2.3 Persistence of cross-modal effects

In a related study of gustatory manipulation by visual modulation using cross-modal effects [4, 14], the foods eaten are often bite-sized foods that can be held in hand, such as cookies and sushi. These reasons are due to the difficulty in accommodating food that decreases in quantity and deforms as it is eaten by the user and the problem of difficulty in eating when the HMD is worn. Please note that some studies address these issues, such as the gustatory manipulation interface [4] that supports the deformation of AR markers printed on cookies. The problem with using bite-sized foods is that it is difficult to verify persist of gustatory manipulation by cross-modal effects because the food is lost quickly.

In the experimental results in Chap. 4 revealed that the gustatory manipulation by visual modulation persisted in many participants. Another exciting aspect of the results is that their persistent multisensory flavor perception changes are divided into three groups: those in which the intensity of the multisensory flavor perception change gradually increased, those in which it gradually decreased, and those in which it did not fluctuate, each with a similar number of participants. The multisensory integration model using Bayes' theorem proposed by Ernst *et al.* obtains the posterior distribution by the product of the likelihood distribution of sensory information and the prior distribution developed in our daily food experience [29]. Repeating experiments in which the effect size of a cross-modal effect such as Chap. 4 is increased or decreased may reveal what distribution the multisensory integration model has. Repeating experiments in which the effect size of the cross-modal effect is increased or decreased, as in Chap. 4, may help to construct a more accurate multisensory integration model.

8.2.4 Visual modulation to the surrounding environment

Experimental results in Chap. 5 comparing meals in virtual and real environments found no statistical difference in the effect of visual modifications to the food's surroundings on taste and flavor. A study that investigated the effects of the VE on eating using CAVE reported that changes in the VE changed the evaluation of the food before tasting but did not affect taste and liking after tasting [150]. In a study that investigated the effects of changes in the VE on eating using HMD, changes in the VE did not affect questionnaire evaluations, such as multisensory flavor perception after tasting the food [151]. On the other hand, some studies have reported that eating in the VE affects multisensory flavor perception. Stelick *et al.* reported a more robust pungency of blue cheese when viewing a 360-degree image of a cow barn [152].

In addition, Bouba/Kiki effect [77] investigated the correspondence between speech and the visual shape of objects. In multisensory flavor perception studies, spiky shapes (Kiki) were associated with sourness and rounded shapes (Bouba) with sweetness in both foods and beverages [78, 79]. Cornelio *et al.* found that rounding the actual sweet food shape in the VE increases the sweetness [16]. Chen *et al.* have made beverages taste sweeter in a rounded virtual space [17]. As described above, the effects of VEs on multisensory flavor perception are complex and require further investigation. On the other hand, most of these studies using food in VEs involve beverages or bite-sized foods, or the foods may not be visible in the VE. Ukemochi, which we have developed in Chap. 5, can improve the ease of eating by making the food in the bowl visible while maintaining the presence of the VE. Our future work is to continue to improve Ukemochi so that it can be used in studies of the impact of VEs on multisensory flavor perception.

8.2.5 Visual modulation at the moment the food is placed in the mouth

We developed an HMD with an increased downward FoV that allows visual modulation at the moment the food is placed in the mouth, when the taste is perceived, to increase the cross-modal effect of visual change on gustation. However, the experimental results in Chap. 6 did not find any statistical difference in the amount of cross-modal effect and in improving the ease of eating, despite the visual modulation to the mouth area. The most likely reason that the increase in downward FoV did not increase these effects is that the user may not look at the food when it is placed in the mouth.

Studies that have investigated the relationship between actual eating and eye gaze include a study that investigated what foods humans look at when choosing food in a buffet environment [153] and a study that revealed that people gaze at their favorite foods just before eating [154]. These studies aim to investigate which foods users view and choose among multiple foods. As far as we could find, we did not find any studies that investigated which parts of the food or foods are looked at during the eating of food. The reason is that commercially available eye trackers do not have a viewing angle that can detect the line of sight near the mouth. There are many ways to eat food, such as holding food in one's hand, eating food with a plate on a table, or eating food with a bowl. Therefore, it is expected that the gazing point during a meal will also vary depending on the type of food. It is also possible that the closer an object is to both eyes, the more difficult it is for humans structurally to see in 3D and that the behavior of gazing at an object near the mouth itself is unnatural. The question of what we are looking at at the moment we taste is a topic that needs to be explored in our investigation of the relationship between vision and gustation, and one that we plan to investigate.

8.2.6 Effects of avatar display and changes in a physical appearance on multisensory flavor perception

We showed in Chap. 7 the positive effects of the first-person view of the avatar's body other than the hands improving presence and SoSL using an HMD with an increased downward FoV. However, due to time limitations, it was impossible to investigate the impact of the avatar display on multisensory flavor perception. Since gustatory studies in the VE have been on the rise recently, we believe that multisensory flavor perception with the application of avatars, which is essential for presence enhancement and multisensory integration, will increase in the future [16, 17, 18, 19, 20]. Whereas we have question remains whether the display of avatars necessarily has a positive effect on multisensory scent perception.

We hypothesize that the use of avatars improves the likelihood of visual information and promotes multisensory integration. The reason is that the closer the appearance of the self-avatar is to one's own body, the higher the SoBO is evoked [23, 98]. However, currently used avatars are limited in appearance and body movement fidelity by low-resolution textures and IK limitations to allow the avatar to follow its movements. Experiments displaying such avatars may be counterproductive to improving visual likelihood.

We also hypothesize that changes in avatar appearance alter multisensory flavor perception. The reason is that related studies have shown that changing the appearance of an avatar can change the user's behavior and psychological state [30, 31, 32]. Although, there is a problem in that the visual likelihood decreases when the avatar is changed to an appearance different from the actual body. We are concerned that this reduction in visual likelihood may adversely affect multisensory flavor perception. We should also be noted that the changes in avatar appearance reported in related studies do not necessarily significantly impact humans. In addition, the experimental results we reported in Chap. 6 indicate that participants may not look at the food displayed in the downward FoV when eating. Without careful experimental design focusing on what humans pay attention to while eating, avatar display, and appearance changes may not affect multisensory flavor perception.

Despite these limitations and problems, combining avatars with multisensory flavor perception is worthwhile. It is expected that changes in the avatar's apparent weight will change the user's perception of their weight [155, 156] and may be applied to support the treatment of background body image disturbance, a symptom of anorexia nervosa [157]. We may be able to contribute to the treatment of body image disturbance by having the patient eat with the avatar's weight altered. The effect of the avatar's change in appearance is expected to be applied to reduce prejudice due to racial discrimination caused by differences in skin color [158, 159, 160]. Experiencing the food culture and multisensory flavor perception of other countries through the effect of the avatar change may reduce prejudice. Adults using child avatars have been reported to think and act more like children [161]. Reliving childhood picky eaters and likes and dislikes may help them understand thier child and develop ways to help them eat foods they don't like. Above all, changing the avatar's appearance to that of a rich person or chef may make the food seem more luxurious and help us appreciate it. Thus, studies combining avatars and multisensory flavor perception can be expected to have various effects.

Related studies that change avatars' appearance uses virtual mirrors installed in the VE. Even if the avatar is not visible from the downward FoV, the avatar may affect multisensory flavor perception if the user can recognize how their appearance is changing. Improving the fidelity of avatars' appearance and behavior is currently being actively researched, and problems caused by these factors may be solved in a few years. We believe that by investigating the conditions under which an avatar displayed in the downward FoV is accepted as oneself and how it is represented, we can apply the effects of changes in the avatar's appearance to the study of multisensory flavor perception without reducing the likelihood of visual information.

8.3 Summary

In this doctoral thesis, we investigated the effects of visual changes on multisensory flavor perception and developed a gustatory manipulation interface that can continuously change the taste and type of food and present smell and food texture. We have also developed a system for constructing a dining experience that enables eating in the VE while maintaining presence and ease of eating, as well as an HMD with an increased downward FoV that can present visual information near the mouth. Although these systems had little effect on multisensory flavor perception, they are necessary exploratory research to investigate the effects of changes in the surrounding environment and the presentation of visual information at the moment of multisensory flavor perception. We also used an HMD with an increased downward FoV to improve presence and SoSL by displaying the avatar's body in first-person perspective, except for the avatar's hands. We created the necessary environment to investigate the effects of avatar display and appearance changes on multisensory flavor perception. Eating food is necessary for survival and an act of longing that can be compared to love. We believe that promoting the study of gustatory manipulation interface can improve our daily lives.

Acknowledgements

I studied at the Nara Institute of Science and Technology (NAIST) for five years in both Master's and Doctoral programs. It seems like a very long period in numbers, but it feels like a moment to me. When I first visited NAIST, the season was late spring, with the anticipation of summer heat. I can still recall the pleasant, crisp energy of the summer sun. It's a pity that I also remember the feeling of carrying a heavy carry-on case up the long hill to the station. As I am finishing my doctoral dissertation, the weather is snowing so hard that it is white in front of my eyes (It was the coldest winter wave in 10 years). However, as a long-time NAIST student, I know that pleasant temperatures and blessed cherry blossoms await us at graduation. My time at NAIST has been very enjoyable and productive for my development as a researcher. At the end of the doctoral dissertation, I would like to thank all people who gave me favors in this period.

I have learned a great deal at my disposal in the Cybernetics and Reality Engineering (CARE) Laboratory. I am deeply grateful to my supervisor Prof. Kiyoshi Kiyokawa. He always carefully guided me on how to sublimate my many ideas into research and how to develop them into deep research. He was also very generous in accepting my very free-spirited behavior. He was willing to accept not only my collaborative research activities outside the university, but also challenges not directly related to research, such as the IPA Mitou project to deepen technology and the Kuma foundation challenge, which specializes in art. It has allowed me to develop outstanding research skills and many more. I would like to show my gratitude to him by continuing to take on these challenges, and by working with freedom.

I want to thank the many co-supervisors who have helped me at NAIST. Prof. Hirokazu Kato, a sub-chief examiner, gave me a lot of valuable suggestions to complete this doctoral thesis. The feedback I received at the interim review meeting and the doctoral dissertation review helped me to consider how I should devote myself as a researcher. Assoc. Prof. Hideaki Uchiyama gave me professional research guidance and many compliments. I learned how to motivate and mentor junior lab members and students I will be mentoring in the future. Asst. Prof. Naoya Isoyama is so young and fresh that I mistook her for a student when I first met him. His attitude toward research and our laboratory management was very helpful to me when I was thinking about my career as a researcher. Assoc. Prof. Nobuchika Sakata, who has transferred to Ryukoku University, has given me many tasty booze tastes. What is more, they provided us with guidance on prototyping methods to nurture the seeds of research and helped us to create a research environment that made this possible.

I want to thank the many co-authors who helped me with my dissertation, journal submissions, and international conference presentations. Assoc. Prof. Takuji Narumi at the University of Tokyo gave me a lot of expert knowledge about gustatory presentation methods and the influence of avatar specifications on users. Assoc. Prof. Takuji Narumi at the University of Tokyo gave me a lot of expert knowledge about gustatory presentation methods and the influence of avatars on users. I am glad he will continue supervising me as a project assistant professor next year. I am grateful to Prof. Keiji Yanai at the University of Electro-Communications, Project Assoc. Prof. Norihiko Kawai at the Osaka Institute of Technology, Dr. Diego Vilela Monteiro at ESIEA, Prof. Tomohiro Amemiya, and Dr. Akimi Oyanagi of the University of Tokyo as co-authors. Without their support, I would not have been able to write my dissertation. I want to thank Daichi Horita, a Ph.D. student at the University of Tokyo. He has generously contributed his knowledge and skills in the field of machine learning to my research on gustatory manipulation methods. I am looking forward to him showing me the best Jiro Ramen restaurants in Tokyo.

I was able to be involved in various extramural projects in the technical and artistic fields as well as in research. I want to thank my friends in the extramural projects and my project manager for introducing me to many enjoyable worlds after being a hermit at NAIST. Prof. Masahiko Inami at the University of Tokyo and the project managers are my mentors at IPA Mitou, and my friends who encouraged each other taught me the skills, know-how, and enthusiasm to turn ideas into products and businesses. At the same time, they also gave us a taste of fine meats, which is necessary for studying multisensory flavor perception. At the Kuma foundation, I got to know creatives my age with a high level of artistry and production skills. They taught me how to express my worldview and enjoy a craft beer taste while we exhibited our work together. The Japan Society for the Promotion of Science Research Fellowship financially supported my life in the doctoral program. The funds I received have literally become my flesh and blood.

During my time at CARE Laboratory, I have been blessed with my lab members. Dr. Daiki Hagimori is a great friend with whom I spent time from when I entered the lab to when I graduated from the Ph.D. program. I initially thought my doctoral journey would be solitary and lonely. I want to thank him because he was the reason I did not falter and was able to complete my long journey with joy and laughter. I also dread to think that my workload would have more than doubled without him. Our excellent secretaries Mina Nakamura and Yoshie Sato, helped me in many aspects of my life in the laboratory. They especially helped me a lot when I missed the flight to go abroad for an international conference. I want to express my deepest apologies and gratitude. Once again, I would like to express my great appreciation to all the lab members. The time I spent in the lab with them was fun and unforgettable.

Finally, I want to thank my parents and sister. They allowed me to embark on the doctoral journey, despite their concerns about my disease. It is my great fortune to be in their family.

Again and again, thank you so much to everyone who has been a part of my doctoral journey over the past five years. I gained many diverse experiences and memories I could not have imagined when I first entered the school. Without one of them, I would not have been where I am today. I will continue to travel and take on more challenges to show my appreciation.

本当にありがとうございました.

Bibliography

[1] Robert P Erickson. “A Study of the Science of Taste: On the Origins and Influence of the Core Ideas”. In: *Behavioral and Brain Sciences* 31.1 (2008), pp. 59–75. DOI: 10.1017/S0140525X08003348.

[2] Charles Spence. “Multisensory Flavor Perception”. In: *Cell* 161.1 (2015), pp. 24–35. DOI: 10.1016/j.cell.2015.03.007.

[3] John Prescott. “Multisensory processes in flavour perception and their influence on food choice”. In: *Current Opinion in Food Science* 3 (2015), pp. 47–52. DOI: 10.1016/j.cofs.2015.02.007.

[4] Takuji Narumi, Shinya Nishizaka, Takashi Kajinami, Tomohiro Tanikawa, and Michitaka Hirose. “Augmented reality flavors: Gustatory display based on Edible Marker and cross-modal interaction”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2011, pp. 93–102. DOI: 10.1145/1978942.1978957.

[5] Naoya Koizumi, Hidekazu Tanaka, Yuji Uema, and Masahiko Inami. “Chewing jockey: augmented food texture by using sound based on the cross-modal effect”. In: *Proceedings of the International Conference on Advances in Computer Entertainment Technology*. 2011, pp. 1–4. DOI: 10.1145/2071423.2071449.

[6] Hiroo Iwata, Hiroaki Yano, Takahiro Uemura, and Tetsuro Moriya. “Food simulator: A haptic interface for biting”. In: *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces*. 2004, pp. 51–57. DOI: 10.1109/VR.2004.1310055.

[7] Massimiliano Zampini, Emma Wantling, Nicola Phillips, and Charles Spence. “Multisensory flavor perception: Assessing the influence of fruit acids and color cues on the perception of fruit-flavored beverages”. In: *Food Quality*

and Preference 19.3 (2008), pp. 335–343. DOI: 10.1016/J.FOODQUAL.2007.11.001.

[8] Gil Morrot, Frédéric Brochet, and Denis Dubourdieu. “The Color of Odors”. In: *Brain and Language* 79.2 (2001), pp. 309–320. DOI: 10.1006/BRLN.2001.2493.

[9] Takuji Narumi, Munehiko Sato, Tomohiro Tanikawa, and Michitaka Hirose. “Evaluating cross-sensory perception of superimposing virtual color onto real drink: toward realization of pseudo-gustatory displays”. In: *Proceedings of the Augmented Human International Conference.* 2010, pp. 18–24. DOI: 10.1145/1785455.1785473.

[10] Nimesha Ranasinghe, Thi Ngoc Tram Nguyen, Yan Liangkun, Lien-Ya Lin, David Tolley, and Ellen Yi-Luen Do. “Vocktail: A Virtual Cocktail for Pairing Digital Taste, Smell, and Color Sensations”. In: *Proceedings of the ACM international conference on Multimedia.* 2017, pp. 1139–1147. DOI: 10.1145/3123266.3123440.

[11] Betina Piqueras-Fiszman, Jorge Alcaide, Elena Roura, and Charles Spence. “Is it the plate or is it the food? Assessing the influence of the color (black or white) and shape of the plate on the perception of the food placed on it”. In: *Food Quality and Preference* 24.1 (2012), pp. 205–208. DOI: 10.1016/J.foodqual.2011.08.011.

[12] Maya Shankar, Christopher Simons, Baba Shiv, Samuel McClure, Carmel A. Levitan, and Charles Spence. “An expectations-based approach to explaining the cross-modal influence of color on orthonasal olfactory identification: The influence of the degree of discrepancy”. In: *Attention, Perception, and Psychophysics* 72.7 (2010), pp. 1981–1993. DOI: 10.3758/APP.72.7.1981.

[13] Elisabetta Làdavas. “Functional and dynamic properties of visual peripersonal space”. In: *Trends in Cognitive Sciences* 6 (1 Jan. 2002), pp. 17–22. DOI: 10.1016/S1364-6613(00)01814-3.

[14] Junya Ueda and Katsunori Okajima. “AR food changer using deep learning and cross-modal effects”. In: *Proceedings of IEEE AIVR.* 2019, pp. 110–117. DOI: 10.1109/AIVR46125.2019.00025.

[15] H. L. Meiselman, J. L. Johnson, W. Reeve, and J. E. Crouch. "Demonstrations of the influence of the eating environment on food acceptance". In: *Appetite* 35.3 (2000), pp. 231–237. DOI: 10.1006/APPE.2000.0360.

[16] Patricia Cornelio, Christopher Dawes, Emanuela Maggioni, et al. "Virtually tasty: An investigation of the effect of ambient lighting and 3D-shaped taste stimuli on taste perception in virtual reality". In: *International Journal of Gastronomy and Food Science* 30 (Dec. 2022), p. 100626. DOI: 10.1016/J.IJGFS.2022.100626.

[17] Yang Chen, Arya Xinran Huang, Ilona Faber, Guido Makransky, and Federico J.A. Perez-Cueto. "Assessing the Influence of Visual-Taste Congruency on Perceived Sweetness and Product Liking in Immersive VR". In: *Foods 2020, Vol. 9, Page 465* 9 (4 Apr. 2020), p. 465. DOI: 10.3390/FOODS9040465.

[18] Dannie Korsgaard, Niels Christian Nilsson, and Thomas Bjørner. "Immersive eating: Evaluating the use of head-mounted displays for mixed reality meal sessions". In: *IEEE 3rd Workshop on Everyday Virtual Reality.* 2017, pp. 1–4. DOI: 10.1109/WEVR.2017.7957709.

[19] Pablo Perez, Ester Gonzalez-Sosa, Redouane Kachach, Jaime Ruiz, Ignacio Benito, Francisco Pereira, and Alvaro Villegas. "Immersive Gastronomic Experience with Distributed Reality". In: *IEEE 5th Workshop on Everyday Virtual Reality.* Institute of Electrical and Electronics Engineers Inc., 2019, pp. 1–6. DOI: 10.1109/WEVR.2019.8809591.

[20] Dannie Korsgaard, Thomas Bjorner, Jon R. Bruun-Pedersen, Pernille K. Sorensen, and Federico J.A. Perez-Cueto. "Eating together while being apart: A pilot study on the effects of mixed-reality conversations and virtual environments on older eaters' solitary meal experience and food intake". In: *Proceedings of the IEEE VR.* 2020, pp. 365–370. DOI: 10.1109/VRW50115.2020.00079.

[21] Andrea Serino, Jean Paul Noel, Robin Mange, et al. "Peripersonal space: An index of multisensory body-environment interactions in real, virtual, and mixed realities". In: *Frontiers in ICT* 4 (JAN Jan. 2017), p. 31. DOI: 10.3389/FICT.2017.00031/BIBTEX.

[22] Jean Paul Noel, Olaf Blanke, and Andrea Serino. "From multisensory integration in peripersonal space to bodily self-consciousness: from statistical regularities to statistical inference". In: *Annals of the New York Academy of Sciences* 1426 (1 Aug. 2018), pp. 146–165. DOI: 10.1111/NYAS.13867.

[23] Konstantina Kilteni, Raphaela Groten, and Mel Slater. "The Sense of Embodiment in virtual reality". In: *Presence: Teleoperators and Virtual Environments* 21.4 (2012), pp. 373–387. DOI: 10.1162/PRES_a_00124.

[24] Bernhard Spanlang, Jean-Marie Normand, David Borland, et al. "How to Build an Embodiment Lab: Achieving Body Representation Illusions in Virtual Reality". In: *Frontiers in Robotics and AI* 1 (2014), p. 9. DOI: 10.3389/frobt.2014.00009.

[25] Nami Ogawa, Takuji Narumi, Hideaki Kuzuoka, and Michitaka Hirose. "Do You Feel Like Passing Through Walls?: Effect of Self-Avatar Appearance on Facilitating Realistic Behavior in Virtual Environments". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2020, pp. 1–14. DOI: 10.1145/3313831.3376562.

[26] Catherine Preston, Benjamin J. Kuper-Smith, and H. Henrik Ehrsson. "Owning the body in the mirror: The effect of visual perspective and mirror view on the full-body illusion". In: *Scientific Reports* 5.1 (2015), p. 18345. DOI: 10.1038/srep18345.

[27] Erin A. McManus, Bobby Bodenheimer, Stephan Streuber, Stephan De La Rosa, Heinrich H. Bülthoff, and Betty J. Mohler. "The influence of avatar (self and character) animations on distance estimation, object interaction and locomotion in immersive virtual environments". In: *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception in Graphics and Visualization*. ACM Press, 2011, pp. 37–44. DOI: 10.1145/2077451.2077458.

[28] Stephan Streuber, Stephan de la Rosa, Laura Trutoiu, Heinrich H. Bülthoff, and Betty J. Mohler. "Does Brief Exposure to a Self-avatar Effect Common Human Behaviors in Immersive Virtual Environments?" In: *Eurographics 2009 - Short Papers*. The Eurographics Association, 2009. DOI: 10.2312/egs.20091042.

[29] Marc O. Ernst. “A Bayesian view on multimodal cue integration”. In: *Human body perception from the inside out: Advances in visual cognition.* (6 2006), pp. 105–131.

[30] Konstantina Kilteni, Ilias Bergstrom, and Mel Slater. “Drumming in immersive virtual reality: The body shapes the way we play”. In: *IEEE Transactions on Visualization and Computer Graphics* 19 (4 Apr. 2013), pp. 597–605. DOI: 10.1109/TVCG.2013.29.

[31] Domna Banakou, Sameer Kishore, and Mel Slater. “Virtually being Einstein results in an improvement in cognitive task performance and a decrease in age bias”. In: *Frontiers in Psychology* 9 (JUN June 2018), p. 917. DOI: 10.3389/FPSYG.2018.00917/BIBTEX.

[32] Nick Yee and Jeremy Bailenson. “The Proteus Effect: The Effect of Transformed Self-Representation on Behavior”. In: *Human Communication Research* 33.3 (2007), pp. 271–290. DOI: 10.1111/j.1468-2958.2007.00299.x.

[33] Sami Damak, Minqing Rong, Keiko Yasumatsu, et al. “Detection of Sweet and Umami Taste in the Absence of Taste Receptor T1r3”. In: *Science* 301.5634 (2003), pp. 850–853. DOI: 10.1126/science.1087155.

[34] “Individual Differences Among Children in Sucrose Detection Thresholds: Relationship With Age, Gender, and Bitter Taste Genotype”. In: *Nursing Research* 65.1 (2016), pp. 3–12. DOI: 10.1097/NNR.0000000000000138.

[35] 横澤 一彦. “統合的認知”. In: 認知科学 21 (3 2014), pp. 295–303. DOI: 10.11225/JCSS.21.295.

[36] Marc O. Ernst and Martin S. Banks. “Humans integrate visual and haptic information in a statistically optimal fashion”. In: *Nature* 415.6870 (2002), pp. 429–433. DOI: 10.1038/415429a.

[37] Paul Rozin. “"Taste-smell confusions" and the duality of the olfactory sense”. In: *Perception & Psychophysics* 31.4 (1982), pp. 397–401. DOI: 10.3758/BF03202667.

[38] Stephen D. Roper and Nirupa Chaudhari. “Taste buds: Cells, signals and synapses”. In: *Nature Reviews Neuroscience* 18.8 (2017), pp. 485–497. DOI: 10.1038/nrn.2017.68.

[39] Richard J. Stevenson, John Prescott, and Robert A. Boakes. “Confusing tastes and smells: How odours can influence the perception of sweet and sour tastes”. In: *Chemical Senses* 24.6 (1999), pp. 627–635. DOI: 10.1093/chemse/24.6.627.

[40] Claire Murphy and William S. Cain. “Taste and Olfaction: Independence vs Interaction”. In: *Physiology and Behavior* 24.3 (1980), pp. 601–605. DOI: 10.1016/0031-9384(80)90257-7.

[41] B. G. Slocombe, D. A. Carmichael, and J. Simner. “Cross-modal Tactile–taste Interactions in Food Evaluations”. In: *Neuropsychologia* 88 (2016), pp. 58–64. DOI: 10.1016/j.neuropsychologia.2015.07.011.

[42] Massimiliano Zampini and Charles Spence. “The Role of Auditory Cues in Modulating The Perceived Crispness and Staleness of Potato Chips”. In: *Journal of Sensory Studies* 19.5 (2004), pp. 347–363. DOI: 10.1111/j.1745-459x.2004.080403.x.

[43] Charles Spence and Maya U. Shankar. “The Influence of Auditory Cues on the Perception of, and Responses to, Food and Drink”. In: *Journal of Sensory Studies* 25.3 (2010), pp. 406–430. DOI: 10.1111/j.1745-459X.2009.00267.x.

[44] Homei Miyashita. “Norimaki synthesizer: Taste display using ion electrophoresis in five gels”. In: *Proceedings of Extended Abstracts of the CHI Conference on Human Factors in Computing Systems.* 2020, pp. 1–6. DOI: 10.1145/3334480.3382984.

[45] Charles Spence, Carmel A. Levitan, Maya U. Shankar, and Massimiliano Zampini. “Does Food Color Influence Taste and Flavor Perception in Humans?” In: *Chemosensory Perception* 3.1 (2010), pp. 68–84. DOI: 10.1007/s12078-010-9067-z.

[46] L.N. van der Laan, D.T.D. de Ridder, M.A. Viergever, and P.A.M. Smeets. "The first taste is always with the eyes: A meta-analysis on the neural correlates of processing visual food cues". In: *NeuroImage* 55.1 (2011), pp. 296–303. DOI: `10.1016/j.neuroimage.2010.11.055`.

[47] Jennifer A Stillman. "Color Influences Flavor Identification in Fruit-flavored Beverages". In: *Journal of Food Science* 58 (4 July 1993), pp. 810–812. DOI: `10.1111/j.1365-2621.1993.tb09364.x`.

[48] Takuji Narumi, Yuki Ban, Takashi Kajinami, Tomohiro Tanikawa, and Michitaka Hirose. "Augmented perception of satiety: controlling food consumption by changing apparent size of food with augmented reality". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 2012, pp. 109–118. DOI: `10.1145/2207676.2207693`.

[49] Oliver Genschow, Leonie Reutner, and Michaela Wänke. "The color red reduces snack food and soft drink intake". In: *Appetite* 58.2 (2012), pp. 699–702. DOI: `10.1016/J.APPET.2011.12.023`.

[50] Brian Wansink, James E. Painter, and Jill North. "Bottomless Bowls: Why Visual Cues of Portion Size May Influence Intake ". In: *Obesity Research* 13.1 (2005), pp. 93–100. DOI: `10.1038/oby.2005.12`.

[51] Sho Sakurai, Takuji Narumi, Yuki Ban, Tomohiro Tanikawa, and Michitaka Hirose. "Affecting Our Perception of Satiety by Changing the Size of Virtual Dishes Displayed with a Tabletop Display". In: *Proceedings of the International Conference on Virtual, Augmented and Mixed Reality.* 2013, pp. 90–99. DOI: `10.1007/978-3-642-39420-1_11`.

[52] Jay A Lieberman and Scott H Sicherer. "Quality of life in food allergy". In: *Current Opinion in Allergy and Clinical Immunology* 11.3 (2011), pp. 236–242. DOI: `10.1097/ACI.0b013e3283464cf0`.

[53] A. DunnGalvin, B. M. J. de BlokFlokstra, A. W. Burks, A. E. J. Dubois, and J. O'B. Hourihane. "Food allergy QoL questionnaire for children aged 0–12 years: content, construct, and cross-cultural validity". In: *Clinical & Experimental Allergy* 38.6 (2008), pp. 977–986. DOI: `10.1111/j.1365-2222.2008.02978.x`.

[54] Yuri Matsubara, Ryusuke Ae, Yukihiro Ohya, et al. “Estimated Number of Patients with Food Allergy in Japan: The Present Status and Issues Regarding Epidemiological Investigation”. In: *Japanese Journal of Allergology* 67.6 (2018), pp. 767–773. DOI: 10.15036/arerugi.67.767.

[55] 林典子、今井孝成、長谷川実穂、黒坂了正、佐藤さくら、小俣貴嗣、富川盛光、宿谷 明紀、海老澤 元宏. “食物アレルギー児と非食物アレルギー児の食生活のQOL （Quality of life）比較調査”. In: 日本小児アレルギー学会誌 23.5 (2009), pp. 643–650.

[56] A. J. Cummings, R. C. Knibb, R. M. King, and J. S. Lucas. “The psychosocial impact of food allergy and food hypersensitivity in children, adolescents and their families: a review”. In: *Allergy* 65.8 (2010), pp. 933–945. DOI: 10.1111/j.1398-9995.2010.02342.x.

[57] Natalie J. Avery, Rosemary M. King, Susan Knight, and Jonathan O’B. Hourihane. “Assessment of quality of life in children with peanut allergy”. In: *Pediatric Allergy and Immunology* 14.5 (2003), pp. 378–382. DOI: 10.1034/j.1399-3038.2003.00072.x.

[58] 豊田恵美子、山崎安信、岡浩一朗. “クローン病患者における口腔関連Quality of Life と口腔保健行動”. In: 口腔衛生学会雑誌 62.3 (2012), pp. 322–328.

[59] Xiaoang Wan, Andy T. Woods, Jasper J. F. van den Bosch, Kirsten J. McKenzie, Carlos Velasco, and Charles Spence. “Cross-cultural differences in crossmodal correspondences between basic tastes and visual features”. In: *Frontiers in Psychology* 5.DEC (2014), p. 1365. DOI: 10.3389/fpsyg.2014.01365.

[60] Xiaoang Wan, Carlos Velasco, Charles Michel, Bingbing Mu, Andy T Woods, and Charles Spence. “Does the type of receptacle influence the crossmodal association between colour and flavour? A cross-cultural comparison”. In: *Flavour* 3.1 (2014), p. 3. DOI: 10.1186/2044-7248-3-3.

[61] Mary K. Ngo, Carlos Velasco, Alejandro Salgado, Emilia Boehm, Daniel O’Neill, and Charles Spence. “Assessing crossmodal correspondences in exotic fruit juices: The case of shape and sound symbolism”. In: *Food Qual-*

ity and Preference 28.1 (2013), pp. 361–369. DOI: 10.1016/J.FOODQUAL.2012.10.004.

[62] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems.* 2014, pp. 2672–2680.

[63] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein Generative Adversarial Networks". In: *Proceedings of the International Conference on Machine Learning.* 2017, pp. 214–223.

[64] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *Proceedings of the International Conference on Learning Representations.* 2016.

[65] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. "Progressive Growing of GANs for Improved Quality, Stability, and Variation". In: *Proceedings of the International Conference on Learning Representations.* 2017.

[66] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. "Image-to-Image Translation with Conditional Adversarial Networks". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2017.

[67] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *Proceedings of the IEEE International Conference on Computer Vision.* 2017, pp. 2242–2251. DOI: 10.1109/ICCV.2017.244.

[68] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. "Learning to Discover Cross-Domain Relations with Generative Adversarial Networks". In: *Proceedings of the International Conference on Machine Learning.* 2017.

[69] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. “Unsupervised Image-to-Image Translation Networks”. In: *Advances in Neural Information Processing Systems*. 2017.

[70] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. “Generative Adversarial Text to Image Synthesis”. In: *Proceedings of the 33rd International Conference on Machine Learning*. 2016.

[71] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. “StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 5908–5916. DOI: 10.1109/ICCV.2017.629.

[72] C. Ledig, L. Theis, F. Huszar, et al. “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2016.

[73] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. “StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation”. In: *Proceedings of IEEE CVPR*. 2018, pp. 8789–8797.

[74] Hajime Hoashi Yuji Matsuda and Keiji Yanai. “Recognition of Multiple-Food Images by Detecting Candidate Regions”. In: *ICME* (2012).

[75] D. Horita, R. Tanno, W. Shimoda, and K. Yanai. “Food Category Transfer with Conditional CycleGAN and a Large-scale Food Image Dataset”. In: *Proc. MADiMa*. 2018.

[76] Mara Dionísio, Duarte Teixeira, Poan Shen, Mario Dinis, Monchu Chen, Nuno Nunes, Valentina Nisi, and José Paiva. “Eat&Travel: A New Immersive Dining Experience for Restaurants”. In: *Lecture Notes in Computer Science* 8253 LNCS (2013), pp. 532–535. DOI: 10.1007/978-3-319-03161-3_46.

[77] Wolfgang Köhler. *Gestalt psychology*. Liveright, 1929.

[78] Carlos Velasco, Andy T. Woods, Ophelia Deroy, and Charles Spence. "Hedonic mediation of the crossmodal correspondence between taste and shape". In: *Food Quality and Preference* 41 (Apr. 2015), pp. 151–158. DOI: 10.1016/J.FOODQUAL.2014.11.010.

[79] Alejandro Salgado-Montejo, Jorge A. Alvarado, Carlos Velasco, Carlos J. Salgado, Kendra Hasse, and Charles Spence. "The sweetest thing: the influence of angularity, symmetry, and the number of elements on shape-valence and shape-taste matches". In: *Frontiers in Psychology* 6 (Sept. 2015), p. 1382. DOI: 10.3389/FPSYG.2015.01382/BIBTEX.

[80] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *CVPR*. IEEE Computer Society, 2015, pp. 431–440. DOI: 10.1109/CVPR.2015.7298965.

[81] Liang Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4 (2018), pp. 834–848. DOI: 10.1109/TPAMI.2017.2699184.

[82] Takumi Ege, Wataru Shimoda, and Keiji Yanai. "A New Large-scale Food Image Segmentation Dataset and Its Application to Food Calorie Estimation Based on Grains of Rice". In: *Proceedings of ACMMM Workshop on Multimedia Assisted Dietary Management*. Association for Computing Machinery, 2019, 82–87. DOI: 10.1145/3347448.3357162.

[83] Kaimu Okamoto and Keiji Yanai. "UEC-FoodPix Complete: A Large-Scale Food Image Segmentation Dataset". In: *Lecture Notes in Computer Science* 12665 LNCS (2021), pp. 647–659. DOI: 10.1007/978-3-030-68821-9_51.

[84] Tal Makovski, Gustavo A. Vázquez, and Yuhong V. Jiang. "Visual Learning in Multiple-Object Tracking". In: *PLOS ONE* 3.5 (5 2008), pp. 1–6. DOI: 10.1371/journal.pone.0002228.

[85] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H.S. Torr. "Fast Online Object Tracking and Segmentation: A Unifying Approach". In: *CVPR*. IEEE Computer Society, 2019, pp. 1328–1338. DOI: 10.48550/arxiv.1812.05050.

[86] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. "ECO: Efficient Convolution Operators for Tracking". In: *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* 2017-January (Nov. 2016), pp. 6931–6939. DOI: 10.48550/arxiv.1611.09224.

[87] James Jeng Weei Lin, Henry B.L. Duh, Donald E. Parker, Habib Abi-Rached, and Thomas A. Furness. "Effects of field of view on presence, enjoyment, memory, and simulator sickness in a virtual environment". In: *Proceedings of the Virtual Reality Annual International Symposium*. 2002, pp. 164–171. DOI: 10.1109/vr.2002.996519.

[88] Joshua Ratcliff, Alexey Supikov, Santiago Alfaro, and Ronald Azuma. "ThinVR: Heterogeneous microlens arrays for compact, 180 degree FOV VR near-eye displays". In: *IEEE Transactions on Visualization and Computer Graphics* 26.5 (2020), pp. 1981–1990. DOI: 10.1109/TVCG.2020.2973064.

[89] Ismo Rakkolainen, Roope Raisamo, Matthew Turk, and Tobias Höllerer. "Field-of-view extension for VR viewers". In: *Proceedings of the International Academic Mindtrek Conference*. Association for Computing Machinery, Inc, 2017, pp. 227–230. DOI: 10.1145/3131085.3131088.

[90] Robert Xiao and Hrvoje Benko. "Augmenting the Field-of-View of Head-Mounted Displays with Sparse Peripheral Displays". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press, 2016, pp. 1221–1232. DOI: 10.1145/2858036.2858212.

[91] Wataru Yamada and Hiroyuki Manabe. "Expanding the Field-of-View of Head-Mounted Displays with Peripheral Blurred Images". In: *Proceedings of the Annual Symposium on User Interface Software and Technology*. ACM Press, 2016, pp. 141–142. DOI: 10.1145/2984751.2985735.

[92] Yoshio Ishiguro and Jun Rekimoto. “Peripheral vision annotation: Noninterference information presentation method for mobile augmented reality”. In: *Proceedings of the Augmented Human International Conference*. ACM Press, 2011, pp. 1–5. DOI: 10.1145/1959826.1959834.

[93] Robert W. Lindeman. “A low-cost, low-latency approach to dynamic immersion in occlusive head-mounted displays”. In: *Proceedings of the Virtual Reality*. IEEE Computer Society, 2016, pp. 221–222. DOI: 10.1109/VR.2016.7504733.

[94] Kien T.P. Tran, Sungchul Jung, Simon Hoermann, and Robert W. Lindeman. “MDI: A Multi-channel Dynamic Immersion Headset for Seamless Switching between Virtual and Real World Activities”. In: *Proceedings of the International Symposium on Mixed and Augmented Reality*. IEEE, 2019, pp. 350–358. DOI: 10.1109/VR.2019.8798240.

[95] Isamu Endo, Kazuki Takashima, Maakito Inoue, Kazuyuki Fujita, Kiyoshi Kiyokawa, and Yoshifumi Kitamura. “A Reconfigurable Mobile Head-Mounted Display Supporting Real World Interactions”. In: *Conference on Human Factors in Computing Systems - Proceedings* (May 2021). DOI: 10.1145/3411763.3451765.

[96] Mel Slater, Daniel Perez-Marcos, H. Henrik Ehrsson, and Maria V. Sanchez-Vives. “Inducing illusory ownership of a virtual body”. In: *Frontiers in Neuroscience* 3.SEP (2009), pp. 214–220. DOI: 10.3389/neuro.01.029.2009.

[97] Lorraine Lin and Sophie Jörg. “Need a hand? How appearance affects the virtual hand illusion”. In: *Proceedings of the ACM Symposium on Applied Perception*. Association for Computing Machinery Inc, 2016, pp. 69–76. DOI: 10.1145/2931002.2931006.

[98] Konstantina Kilteni, Antonella Maselli, Konrad P. Kording, and Mel Slater. “Over my fake body: Body ownership illusions for studying the multisensory basis of own-body perception”. In: *Frontiers in Human Neuroscience* 9.MAR (2015), p. 141. DOI: 10.3389/fnhum.2015.00141.

[99] Ferran Argelaguet, Ludovic Hoyet, Michaël Trico, and Anatole Lécuyer. "The role of interaction in virtual embodiment: Effects of the virtual hand representation". In: *Proceedings of the Virtual Reality*. IEEE Computer Society, 2016, pp. 3–10. DOI: 10.1109/VR.2016.7504682.

[100] Valentin Schwind, Lorraine Lin, Massimiliano Di Luca, Sophie Jörg, and James Hillis. "Touch with foreign hands: The effect of virtual hand appearance on visual-haptic integration". In: *Proceedings of the ACM Symposium on Applied Perception*. Association for Computing Machinery Inc, 2018, pp. 1–8. DOI: 10.1145/3225153.3225158.

[101] Nami Ogawa, Takuji Narumi, and Michitaka Hirose. "Virtual hand realism affects object size perception in body-based scaling". In: *Proceedings of the Virtual Reality*. Institute of Electrical and Electronics Engineers Inc., 2019, pp. 519–528. DOI: 10.1109/VR.2019.8798040.

[102] Ye Yuan and Anthony Steed. "Is the rubber hand illusion induced by immersive virtual reality?" In: *Proceedings of the Virtual Reality*. 2010, pp. 95–102. DOI: 10.1109/VR.2010.5444807.

[103] Thomas Waltemate, Dominik Gall, Daniel Roth, Mario Botsch, and Marc Erich Latoschik. "The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response". In: *IEEE Transactions on Visualization and Computer Graphics* 24 (4 Apr. 2018), pp. 1643–1652. DOI: 10.1109/TVCG.2018.2794629.

[104] Jean Luc Lugrin, Maximilian Ertl, Philipp Krop, et al. "Any 'Body' There? Avatar Visibility Effects in a Virtual Reality Game". In: *Proceedings of the Virtual Reality*. Institute of Electrical and Electronics Engineers Inc., 2018, pp. 17–24. DOI: 10.1109/VR.2018.8446229.

[105] G. Tieri, E. Tidoni, E. F. Pavone, and S. M. Aglioti. "Mere observation of body discontinuity affects perceived ownership and vicarious agency over a virtual hand". In: *Experimental Brain Research* 233.4 (2015), pp. 1247–1259. DOI: 10.1007/s00221-015-4202-3.

[106] Gaetano Tieri, Emmanuele Tidoni, Enea Francesco Pavone, and Salvatore Maria Aglioti. “Body visual discontinuity affects feeling of ownership and skin conductance responses”. In: *Scientific Reports* 5.1 (2015), p. 17139. DOI: 10.1038/srep17139.

[107] Ye Pan and Anthony Steed. “How Foot Tracking Matters: The Impact of an Animated Self-Avatar on Interaction, Embodiment and Presence in Shared Virtual Environments”. In: *Frontiers in Robotics and AI* 6 (2019), p. 104. DOI: 10.3389/frobt.2019.00104.

[108] J. Adam Jones, David M. Krum, and Mark T. Bolas. “Vertical field-of-view extension and walking characteristics in head-worn virtual environments”. In: *ACM Transactions on Applied Perception* 14.2 (2016), pp. 1–17. DOI: 10.1145/2983631.

[109] Elham Ebrahimi, Leah S. Hartman, Andrew Robb, Christopher C. Pagano, and Sabarish V. Babu. “Investigating the Effects of Anthropomorphic Fidelity of Self-Avatars on Near Field Depth Perception in Immersive Virtual Environments”. In: *Proceedings of the Virtual Reality.* Institute of Electrical and Electronics Engineers Inc., 2018, pp. 1–8. DOI: 10.1109/VR.2018.8446539.

[110] Brian Ries, Victoria Interrante, Michael Kaeding, and Lee Anderson. “The effect of self-embodiment on distance perception in immersive virtual environments”. In: *Proceedings of the ACM Symposium on Virtual Reality Software and Technology.* ACM Press, 2008, pp. 167–170. DOI: 10.1145/1450579.1450614.

[111] Brian Ries, Victoria Interrante, Michael Kaeding, and Lane Phillips. “Analyzing the effect of a virtual avatar’s geometric and motion fidelity on egocentric spatial perception in immersive virtual environments”. In: *Proceedings of the ACM Symposium on Virtual Reality Software and Technology.* ACM Press, 2009, pp. 59–66. DOI: 10.1145/1643928.1643943.

[112] Diane Dewez, Ludovic Hoyet, Anatole Lecuyer, and Ferran Argelaguet. “Studying the Inter-Relation between Locomotion Techniques and Embodiment in Virtual Reality”. In: *Proceedings of the International Symposium on Mixed and Augmented Reality.* Institute of Electrical and Electronics

Engineers Inc., 2020, pp. 452–461. DOI: 10.1109/ISMAR50242.2020.00070.

[113] Daniel S. Marigold and Aftab E. Patla. “Visual information from the lower visual field is important for walking across multi-surface terrain”. In: *Experimental Brain Research* 188.1 (2008), pp. 23–31. DOI: 10.1007/s00221-008-1335-7.

[114] Shirin E. Hassan, John C. Hicks, Hao Lei, and Kathleen A. Turano. “What is the minimum field of view required for efficient navigation?” In: *Vision Research* 47.16 (2007), pp. 2115–2123. DOI: 10.1016/j.visres.2007.03.012.

[115] Veronica Miyasike-daSilva, Jonathan C. Singer, and William E. McIlroy. “A role for the lower visual field information in stair climbing”. In: *Gait and Posture* 70 (2019), pp. 162–167. DOI: 10.1016/j.gaitpost.2019.02.033.

[116] Ferran Argelaguet Sanz, Anne Helene Olivier, Gerd Bruder, Julien Pettre, and Anatole Lecuyer. “Virtual proxemics: Locomotion in the presence of obstacles in large immersive projection environments”. In: *Proceedings of the Virtual Reality*. Institute of Electrical and Electronics Engineers Inc., 2015, pp. 75–80. DOI: 10.1109/VR.2015.7223327.

[117] M. S. Spetter, P. A. M. Smeets, C. de Graaf, and M. A. Viergever. “Representation of Sweet and Salty Taste Intensity in the Brain”. In: *Chemical Senses* 35.9 (2010), pp. 831–840. DOI: 10.1093/chemse/bjq093.

[118] Kizashi Nakano, Daichi Horita, Nobuchika Sakata, Kiyoshi Kiyokawa, Keiji Yanai, and Takuji Narumi. “DeepTaste: Augmented reality gustatory manipulation with GAN-based real-time food-to-food translation”. In: *Proceedings of the International Symposium on Mixed and Augmented Reality*. Institute of Electrical and Electronics Engineers Inc., 2019, pp. 212–223. DOI: 10.1109/ISMAR.2019.000-1.

[119] Randy Pausch, Thomas Crea, and Matthew Conway. “A Literature Survey for Virtual Environments: Military Flight Simulator Visual Systems and Simulator Sickness”. In: *Presence: Teleoperators and Virtual Environments* 1.3 (1992), pp. 344–363. DOI: 10.1162/pres.1992.1.3.344.

[120] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. “Improved Training of Wasserstein GANs”. In: *arXiv preprint arXiv:1704.00028* (2017).

[121] K. Simonyan and A. Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *Proceedings of the International Conference on Learning Representations.* 2015.

[122] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2016, pp. 770–778.

[123] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. *Instance Normalization: The Missing Ingredient for Fast Stylization.* 2016.

[124] Ting Chun Wang, Ming Yu Liu, Jun Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. “High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2018, pp. 8798–8807. DOI: 10.1109/CVPR.2018.00917.

[125] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. “Video-to-Video Synthesis”. In: *Conference on Neural Information Processing Systems.* 2018.

[126] Peter Kim, Jason Orlosky, and Kiyoshi Kiyokawa. “AR Timewarping: A Temporal Synchronization Framework for Real-Time Sensor Fusion in Head-Mounted Displays”. In: *Proceedings of the 9th Augmented Human International Conference.* ACM Press, 2018, pp. 1–8. DOI: 10.1145/3174910.3174919.

[127] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. “The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedures”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM Press, 2011, pp. 143–146. DOI: 10.1145/1978942.1978963.

[128] Maria V. Sanchez-Vives and Mel Slater. “From presence to consciousness through virtual reality”. In: *Nature Reviews Neuroscience* 6.4 (2005), pp. 332–339. DOI: 10.1038/nrn1651.

[129] Lik-Hang Lee, Tristan Braud, Pengyuan Zhou, et al. “All One Needs to Know about Metaverse: A Complete Survey on Technological Singularity, Virtual Ecosystem, and Research Agenda”. In: (2021). arXiv: 2110.05352.

[130] Matej Kristan, Aleš Leonardis, Jiří Matas, et al. “The Sixth Visual Object Tracking VOT2018 Challenge Results”. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 11129 LNCS (2019), pp. 3–53. DOI: 10.1007/978-3-030-11009-3_1.

[131] Joseph Redmon and Ali Farhadi. “YOLOv3: An Incremental Improvement”. In: (Apr. 2018). arXiv: 1804.02767.

[132] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai. “Recognition of Multiple-Food Images by Detecting Candidate Regions”. In: *IEEE International Conference on Multimedia and Expo*. 2012, pp. 25–30. DOI: 10.1109/ICME.2012.157.

[133] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. “The experience of presence: Factor analytic insights”. In: *Presence: Teleoperators and Virtual Environments* 10.3 (2001), pp. 266–281. DOI: 10.1162/105474601300343603.

[134] Mary Hannan-Jones and Sandra Capra. “Developing a valid meal assessment tool for hospital patients”. In: *Appetite* 108 (2017), pp. 68–73. DOI: 10.1016/J.APPET.2016.09.025.

[135] Kizashi Nakano, Naoya Isoyama, Diego Monteiro, Nobuchika Sakata, Kiyoshi Kiyokawa, and Takuji Narumi. “Head-Mounted Display with Increased Downward Field of View Improves Presence and Sense of Self-Location”. In: *IEEE Transactions on Visualization and Computer Graphics* 27.11 (2021), pp. 4204–4214. DOI: 10.1109/TVCG.2021.3106513.

[136] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. “Image Style Transfer Using Convolutional Neural Networks”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 2016-Decem. IEEE Computer Society, 2016, pp. 2414–2423. DOI: `10.1109/CVPR.2016.265`.

[137] Kohei Kanamori, Nobuchika Sakata, Tomu Tominaga, Yoshinori Hijikata, Kensuke Harada, and Kiyoshi Kiyokawa. “Obstacle Avoidance Method in Real Space for Virtual Reality Immersion”. In: *IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2018, pp. 80–89. DOI: `10.1109/ISMAR.2018.00033`.

[138] Holger Regenbrecht and Thomas Schubert. “Measuring Presence in Augmented Reality Environments: Design and a First Test of a Questionnaire”. In: (2021). arXiv: `2103.02831`.

[139] Dewen Cheng, Yongtian Wang, Hong Hua, and Jose Sasian. “Design of a wide-angle, lightweight head-mounted display using free-form optics tiling”. In: *Optics Letters* 36.11 (2011), p. 2098. DOI: `10.1364/ol.36.002098`.

[140] Kizashi Nakano, Daichi Horita, Norihiko Kawai, Naoya Isoyama, Nobuchika Sakata, Kiyoshi Kiyokawa, Keiji Yanai, and Takuji Narumi. “A Study on Persistence of GAN-Based Vision-Induced Gustatory Manipulation”. In: *Electronics* 10.10 (2021), p. 1157. DOI: `10.3390/ELECTRONICS10101157`.

[141] Kizashi Nakano, Daichi Horita, Naoya Isoyama, Hideaki Uchiyama, and Kiyoshi Kiyokawa. “Ukemochi: A Video See-through Food Overlay System for Eating Experience in the Metaverse”. In: *Proceedings of the CHI EA*. 2022, pp. 1–8. DOI: `10.1145/3491101.3519779`.

[142] Kizashi Nakano, Monica Perusquia-Hernandez, Naoya Isoyama, Hideki Uchiyama, and Kiyoshi Kiyokawa. “The Impact of a Head-Mounted Display with an Increased Downward Field of View on Ease of Eating and Cross-Modal Effects”. In: *第27回日本バーチャルリアリティ学会大会論文集*. Sept. 2022, 3E5-2.

[143] Mar Gonzalez-Franco, Eyal Ofek, Ye Pan, et al. “The Rocketbox Library and the Utility of Freely Available Rigged Avatars”. In: *Frontiers in Virtual Reality* 1 (2020), p. 20. DOI: `10.3389/frvir.2020.561558`.

[144] Jean Luc Lugrin, Johanna Latt, and Marc Erich Latoschik. "Avatar anthropomorphism and illusion of body ownership in VR". In: *Proceedings of the Virtual Reality*. Institute of Electrical and Electronics Engineers Inc., 2015, pp. 229–230. DOI: 10.1109/VR.2015.7223379.

[145] Robert S. Kennedy, Norman E. Lane, Kevin S. Berbaum, and Michael G. Lilienthal. "Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness". In: *The International Journal of Aviation Psychology* 3.3 (1993), pp. 203–220. DOI: 10.1207/s15327108ijap0303_3.

[146] William Steptoe, Anthony Steed, and Mel Slater. "Human tails: Ownership and control of extended humanoid avatars". In: *IEEE Transactions on Visualization and Computer Graphics* 19.4 (2013), pp. 583–590. DOI: 10.1109/TVCG.2013.32.

[147] Thomas A. Stoffregen. "Flow Structure Versus Retinal Location in the Optical Control of Stance". In: *Journal of Experimental Psychology: Human Perception and Performance* 11.5 (1985), pp. 554–565. DOI: 10.1037/0096-1523.11.5.554.

[148] Thomas A. Stoffregen and L. James Smart. "Postural instability precedes motion sickness". In: *Brain Research Bulletin* 47.5 (1998), pp. 437–448. DOI: 10.1016/S0361-9230(98)00102-6.

[149] Tabitha C. Peck and Mar Gonzalez-Franco. "Avatar Embodiment. A Standardized Questionnaire". In: *Frontiers in Virtual Reality* 1 (2021), p. 44. DOI: 10.3389/frvir.2020.575943.

[150] Kyösti Pennanen, Johanna Närväinen, Saara Vanhatalo, Roope Raisamo, and Nesli Sozer. "Effect of virtual eating environment on consumers' evaluations of healthy and unhealthy snacks". In: *Food Quality and Preference* 82 (June 2020), p. 103871. DOI: 10.1016/J.FOODQUAL.2020.103871.

[151] James H. Oliver and James H. Hollis. "Virtual Reality as a Tool to Study the Influence of the Eating Environment on Eating Behavior: A Feasibility Study". In: *Foods 2021, Vol. 10, Page 89* 10 (1 Jan. 2021), p. 89. DOI: 10.3390/FOODS10010089.

[152] Alina Stelick, Alexandra G. Penano, Alden C. Riak, and Robin Dando. "Dynamic Context Sensory Testing–A Proof of Concept Study Bringing Virtual Reality to the Sensory Booth". In: *Journal of Food Science* 83 (8 Aug. 2018), pp. 2047–2051. DOI: 10.1111/1750-3841.14275.

[153] Eunice Wang, Yusuf O. Cakmak, and Mei Peng. "Eating with eyes – Comparing eye movements and food choices between overweight and lean individuals in a real-life buffet setting". In: *Appetite* 125 (June 2018), pp. 152–159. DOI: 10.1016/J.APPET.2018.02.003.

[154] Yuka Yasui, Junko Tanaka, Masaki Kakudo, and Masahiro Tanaka. "Relationship between preference and gaze in modified food using eye tracker". In: *Journal of Prosthodontic Research* 63 (2 Apr. 2019), pp. 210–215. DOI: 10.1016/J.JPOR.2018.11.011.

[155] Erik Wolf, Nina Dollinger, David Mal, Carolin Wienrich, Mario Botsch, and Marc Erich Latoschik. "Body Weight Perception of Females using Photorealistic Avatars in Virtual and Augmented Reality". In: *Proceedings - 2020 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2020* (Nov. 2020), pp. 462–473. DOI: 10.1109/ISMAR50242.2020.00071.

[156] Erik Wolf, Marie Luisa Fiedler, Nina Dollinger, Carolin Wienrich, and Marc Erich Latoschik. "Exploring Presence, Avatar Embodiment, and Body Perception with a Holographic Augmented Reality Mirror". In: *Proceedings - 2022 IEEE Conference on Virtual Reality and 3D User Interfaces, VR 2022* (2022), pp. 350–359. DOI: 10.1109/VR51125.2022.00054.

[157] S. C. Mölbert, A. Thaler, B. J. Mohler, et al. "Assessing body image in anorexia nervosa using biometric self-avatars in virtual reality: Attitudinal components rather than visual body size estimation are distorted". In: *Psychological medicine* 48 (4 Mar. 2018), pp. 642–653. DOI: 10.1017/S0033291717002008.

[158] Nahal Norouzi, Kangsoo Kim, Gerd Bruder, Austin Erickson, Zubin Choudhary, Yifan Li, and Greg Welch. "A Systematic Literature Review of Embodied Augmented Reality Agents in Head-Mounted Display Environments". In: *Proceedings of the International Conference on Artificial Re-*

ality and Telexistence Eurographics Symposium on Virtual Environments (2020), pp. 101–111. DOI: 10.2312/EGVE.20201264.

[159] Tabitha C. Peck, Jessica J. Good, and Katharina Seitz. “Evidence of Racial Bias Using Immersive Virtual Reality: Analysis of Head and Hand Motions during Shooting Decisions”. In: *IEEE Transactions on Visualization and Computer Graphics* 27 (5 May 2021), pp. 2502–2512. DOI: 10.1109/TVCG.2021.3067767.

[160] Tabitha C. Peck, Jessica J. Good, Austin Erickson, Isaac Bynum, and Gerd Bruder. “Effects of Transparency on Perceived Humanness: Implications for Rendering Skin Tones Using Optical See-Through Displays”. In: *IEEE Transactions on Visualization and Computer Graphics* 28 (5 May 2022), pp. 2179–2189. DOI: 10.1109/TVCG.2022.3150521.

[161] Domna Banakou, Raphaela Groten, and Mel Slater. “Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110 (31 July 2013), pp. 12846–12851. DOI: 10.1073/PNAS.1306779110/SUPPL_FILE/SM01.MP4.

Publication List

Peer reviewed journal paper

[1] Kizashi Nakano, Daichi Horita, Norihiko Kawai, Nobuchika Sakata, Kiyoshi Kiyokawa, Keiji Yanai, and Takuji Narumi, A Study on Persistence of GAN-Based Vision-Induced Gustatory Manipulation, Electronics, 10(10), p.1157, 2021

[2] Kizashi Nakano, Naoya Isoyama, Diego Monteiro, Nobuchika Sakata, Kiyoshi Kiyokawa, and Takuji Narumi, Head-Mounted Display with Increased Downward Field of View Improves Presence and Sense of Self-Location, IEEE Transactions on Visualization and Computer Graphics, 27(11), pp.4204–4214, 2021

Peer reviewed international conference

[3] Kizashi Nakano, Daichi Horita, Nobuchika Sakata, Kiyoshi Kiyokawa, Keiji Yanai, and Takuji Narumi, Enchanting Your Noodles: GAN-based Real-time Food-to-Food Translation and Its Impact on Vision-induced Gustatory Manipulation, Proceedings of the IEEE Conference on Virtual Reality, pp.1096–1097, 2019

[4] Kizashi Nakano, Daichi Horita, Nobuchika Sakata, Kiyoshi Kiyokawa, Keiji Yanai, and Takuji Narumi, DeepTaste: Augmented Reality Gustatory Manipulation with GAN-based Real-time Food-to-Food Translation, Proceedings of the IEEE International Symposium on Mixed and Augmented Reality, pp.328–339, 2019

[5] Kizashi Nakano, Daichi Horita, Naoya Isoyama, Hideki Uchiyama, and Kiyoshi Kiyokawa, Ukemochi: A Video See-through Food Overlay System for Eating Experience in the Metaverse, CHI Conference on Human Factors in Computing Systems Extended Abstracts, pp.1–8, 2022

Not peer reviewed international conference

[6] Kizashi Nakano, Enchanting Your Noodles: A Gustatory Manipulation Interface by Using GAN-based Real-time Food-to-Food Translation, Workshop on The Future of Computing & Food, 2019

Peer reviewed domestic conference

[7] 中野 萌士, 磯山 直也, 酒田 信親, 清川 清, 下方視野を拡大したHMDの

開発と評価, 一般社団法人情報処理学会シンポジウムインタラクション, pp.106–114, 2020

Not peer reviewed international conference

[8] 中野 萌士, 鳴海 拓志, 酒田 信親, 清川 清, 麺類を対象とした視覚変調による味覚操作インタフェースの有効性評価, 日本VR学会大会論文集, 21B-1, 2018

[9] Riku Otono, Yusuke Shikanai, Kizashi Nakano，Naoya Isoyama，Hideki Uchiyama，and Kiyoshi Kiyokawa, The Proteus Effect in Augmented Reality: Impact of Avatar Age and User Perspective on Walking Behaviors, 第26回日本VR学会大会論文集，1C3-1，2021

[10] 小柳 陽光, 中野 萌士, 鳴海 拓志, 雨宮 智宏, VR体験でのドラゴンの肉の食事が味の認知に及ぼす影響, 第26回日本VR学会大会論文集, 1B1-8, 2021

[11] 藤澤 岳瞭, 中野 萌士, モニカ ペルスキア エルナンデス, 磯山 直也, 内山 英昭, 清川 清, VR体験向上のためのビデオシースルーARと物理ドアを用いた実環境とVR環境間の遷移手法の提案, メディアエクスペリエンス・バーチャル環境基礎研究会, 122(175), pp.1-2, 2022

[12] 中野 萌士, モニカ ペルスキア エルナンデス, 磯山 直也, 内山 英昭, 清川 清, 下方視野を拡大したヘッドマウントディスプレイが食べやすさやクロスモーダル効果に与える影響, 第27回日本VR学会大会論文集, 2G-25, 2022

Patent (filed)

[13] 中野 萌士, 磯山 直也, 酒田 信親, 清川 清, 垂直視野拡大ヘッドマウントディスプレイ装置, 特開2021-140047, 2021

Award

[14] 日本学術支援機構 特別研究員(DC1), VR技術を用いた食物種類の認知操作による味覚知覚メカニズムの解明, 2020

[15] 2019年度未踏IT人材発掘・育成事業スーパークリエータ, `https://www.meti.go.jp/press/2020/05/20200528003/20200528003.html`, 2020

[16] 令和2年度奈良先端科学技術大学院大学優秀学生, 2020

[17] 第一回XR創作大賞Hi-Noguchi賞, https://sites.google.com/view/xr-sousaku/%E3%83%9B%E3%83%BC%E3%83%A0, 2020

[18] ISMAR 2021 Best Student-Authored Journal Paper, 2021

[19] 第18回 IEEE 関西支部 学生研究奨励賞, 2022

Miscellaneous

Media

[20] Web media, Seamless, 「蒲焼さん太郎」にVRを使用して“うな重”を降臨させる!? 匂い以外ほぼ再現された“VRうな重”がとっても美味しそう, https://originalnews.nico/117207, 2018

[21] Web media, Seamless, 奈良先端科学技術大学院大学など、そうめんをラーメンに錯覚させるARとGANを組み合わせたリアルタイム味覚操作システムを発表。白ご飯が焼飯にも, https://shiropen.com/seamless/enchanting-your-noodles, 2019

[22] Radio, J-WAVE STEP ONE, BEHIND THE SCENE, https://twitter.com/stepone813/status/1113253365197066242, 2019

[23] Television, 日本テレビ, news zero, https://www.facebook.com/newszero/posts/2751982091540586, 2019

[24] Web media, マンバ通信, 第1回　VRでドラゴンは食べられるようになるか？『ダンジョン飯』, https://manba.co.jp/manba_magazines/9199, 2019

[25] Television, NHK, 所さん！大変ですよ, 2020

[26] Television, ABCテレビ, ビーバップハイヒール, https://www.asahi.co.jp/be-bop/, 2020

[27] Teaching materials, ベネッセ, チャレンジ小学五年生, 未来発見BOOK6月号, p.9, 2020

[28] Web media, クマ財団, 現実を超えるものを創りたい, https://kuma-foundation.org/news/3194/, 2020

[29] Television, 読売テレビ, ほんわかテレビ, https://www.ytv.co.jp/honwaka/contents/20201218.html, 2020

[30] Teaching materials, ベネッセ, チャレンジ小学五年生, 未来発見BOOK6月号, p.9, 2021

[31] Web media, Seamless, バーチャル内で“本当の食事”を体験できるか？　奈良先端大と東大が検証, https://www.itmedia.co.jp/news/articles/2205/09/news020.html, 2021

[32] Teaching materials, ベネッセ, チャレンジ小学五年生, 未来発見BOOK6月号, p.9, 2022

[33] Web media, 日経BP, 日経クロステック, 「ラーメンを食べた気になる」、ARで見た目を変える未来の食事, https://xtech.nikkei.com/atcl/nxt/column/18/02118/071200004/, 2022

[34] Television, NHK, ヒューマニエンス40億年のたくらみ, https://www.nhk.jp/p/ts/X4VK5R2LR1/episode/te/1ZV9JZPK92/, 2022

[35] Newspaper, 読売新聞, 2023年01月11日朝刊, 2023

Demonstration

[36] NT金沢2018, https://wiki.nicotech.jp/nico_tech/index.php?NT%E9%87%91%E6%B2%A22018, 2018

[37] イノベーション・ジャパン2019, そうめんをラーメンに変化させる味覚操作ARシステム, 2019

[38] KUMA EXHIBITION 2021, Observable Virtually, https://kuma-foundation.org/exhibition2021/, 2021

Lecture

[39] SINAPS関東支部第6回Jamboree(カフェアカデミー)招待講演, https://www.sinaps.or.jp/report/kanto-jamboree6/, 2018

[40] 日本学術会議, SCIENCE CAFE, 私たちが認識する世界を変化させるバーチャル・リアリティ, https://www.scj.go.jp/ja/event/pdf2/200123.pdf, 2019

[41] サポーターズ, 技育展, 無駄開発：Ukemochi, https://docs.google.com/document/d/1FQlAszGRyBlUsI81ke19OHxImmUyK-IsMMTeijzrV2c, 2020

[42] 先端VRドクトラルシンポジウム, 視覚変調による味覚操作に関する研究, https://vr.u-tokyo.ac.jp/doctoral_symposium_files/, 2021

[43] バーチャル学会2021, オーガナイズドセッション VR × 食, https://sites.google.com/view/virtualconference-2021, 2021

[44] FIT2022 第21回情報科学技術フォーラム，垂直下方向の視野を拡大したヘッドマウントディスプレイはプレゼンスと自己位置感覚を向上させる, https://onsite.gakkai-web.net/fit2022/abstract/data/html/event/event_TCS7-2.html, 2022

Academic conference management

[45] 第8回サイエンス・インカレ, 2018

[46] 第24回日本バーチャルリアリティ学会大会, ギラギラ夏祭り2019 融けろVR, https://conference.vrsj.org/ac2019/program/common/doc/pdf/7C.pdf, 2019

[47] バーチャル学会2019, https://sites.google.com/view/virtualconference-2019, 2019

[48] 第25回日本バーチャルリアリティ学会大会, ギラギラ夏祭り2020 〜ゆるふわ交流会〜, https://connpass.com/event/187508/, 2020

[49] バーチャル学会2020, https://sites.google.com/view/virtualconference2020, 2020

Funds

[50] IPA未踏事業2019, VR空間における食体験の構築, https://www.ipa.go.jp/jinzai/mitou/2019/gaiyou_in-2.html, 2,304,000円, 2018

[51] KAKENHI, 20J21546, VR技術を用いた食物種類の認知操作による味覚知覚メカニズムの解明, 2,500,000円, 2020

[52] クマ財団, クリエータ奨学金, https://kuma-foundation.org/student/kizashi-nakano/, 1,200,000円, 2020

Conference Participation Report

[53] 日本バーチャル学会, VRSJ Newsletter 2019年12月号 (Vol.24, No.12), ISMAR2019, https://vrsj.org/report/10862/, 2019

[54] 光産業技術振興協会, 国際会議速報 2020-No.16 IEEE ISMAR 2020ショート速報, 2020

[55] 日本バーチャル学会, VRSJ Newsletter 2022年5月号 (Vol.27, No.5), CHI2022, https://vrsj.org/report/11836/, 2022