

VISTURE: A System for Video-Based Gesture and Speech Generation by Robots

Kaon Shimoyama
shimoyama@ailab.ics.keio.ac.jp
Keio University
Kohoku, Yokohama, Kanagawa, Japan

Mitsuhiko Kimoto
kimoto@atr.jp
ATR/Keio University
Seika, Soraku, Kyoto, Japan

Kohei Okuoka
okuoka@ailab.ics.keio.ac.jp
Keio University
Kohoku, Yokohama, Kanagawa, Japan

Michita Imai
michita@keio.jp
Keio University
Kohoku, Yokohama, Kanagawa, Japan

ABSTRACT

This paper proposes VISTURE, a system for generating a robot's gesture and speech by using video as input. VISTURE assumes a situation in which a robot conveys what it saw with a camera to a person who was absent. The value of this paper is that we have performed a case study to investigate the expressions that Japanese people use to describe video scenes, and used the results to build VISTURE. In particular, we found classification of expressions depicting the video scenes throughout the case study: Foreground information that is the relevant event of the scene and Background one that is not the main point of the description giving the entire scene. Foreground and Background are referred in combination. VISTURE employs the classification to generate human-like expressions. Moreover, we designed the method to determine Foreground and Background, and it can generate multiple combinations of expressions. We investigated the people's impression of a robot performing the gestures and speech generated by VISTURE to evaluate the quality of those gestures and speech. The results showed that the robot was perceived as more likable and capable when it performed gestures.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

Human-robot interaction, Gesture generation, Speech generation

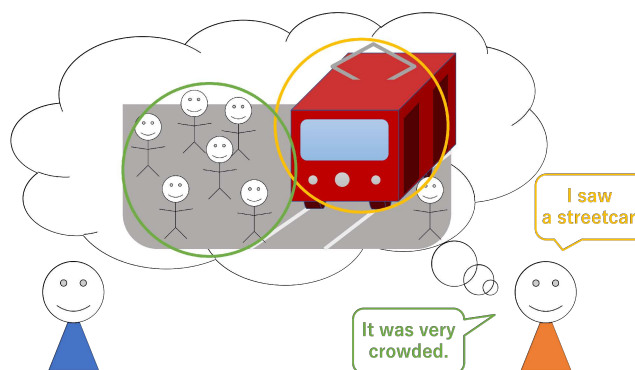


Figure 1: The human on the right is imaging the situation that several people are walking and running near a streetcar. If he focuses on the crowd (green circle), he may say “It was very crowded.” If he focuses on the streetcar (yellow circle), he may say “I saw a streetcar.”

1 INTRODUCTION

Humans commonly communicate something they have seen to people in multimodal ways including speech, gaze, and gestures. In this paper, we focus on a robot which uses video acquired from its own camera and conveys what it saw in a multimodal way.

There are many previous works that use gestures to convey what it sees. ([12, 19]). Nihei et al. proposed a method to generate gestures that represent objects' shape by using images as input [19]. For text generation using video, there is video captioning ([7, 10, 15, 26]). Donahue et al. proposed generating captions that express the input videos by using long short-term memory (LSTM) [7].

However, previous works do not take into account information selection. When humans conveys what they saw to someone who did not see, they are inherently capable of selecting information. For example, in a scene where a streetcar is running through a crowd, as illustrated in Figure 1, whether to focus on the crowd or the streetcar, or whether to focus on other aspects of the scene, will depend on the person and the situation. In particular, if the place is famous for being always crowded, humans would not mention the crowds, but if it is not, humans would mention the crowds. This needs to be taken into account when generating robot representations.

In this study, we consider a system that selects information and generates gestures and speech with reference to human representations, assuming a situation in which a robot conveys what it saw to someone who did not see it. We performed a case study to investigate the expressions that people use to describe video to someone who don't see it, and used the results to build VISTURE. VISTURE is a system for generating a robot's gesture and speech by using video as input, assuming a situation in which a robot conveys what it saw with a camera to a person who was absent. VISTURE can make human-like expressions inspired by humans' expressions. In particular, we found classification of expressions depicting the video scenes throughout the case study: Foreground information that is the relevant event of the scene and Background one that is not the main point of the description giving the entire scene (See Table 2 for details). Foreground and Background are referred in combination. Moreover, we designed the method to determine objects to be mentioned as Foreground and Background, and it can generate multiple combinations of expressions. VISTURE computes features based on the motions of objects in the video and then decides which objects to represent.

The rest of this paper is organized as follows. Section 2 introduces related research. Section 3 describes the target system and the data collection, and Section 4 explains the details of VISTURE. Section 5 details the experiments, and Section 6 presents the results of the experiments. Section 7 discusses the results and Section 8 concludes the paper.

2 RELATED WORK

2.1 Gestures Explaining Scenes

Gestures are important in communication ([3, 5, 11, 13, 20, 23]). Cabibihan et al. showed that humans can understand spatial locations by using ambiguous explanations combined with pointing gestures, as well as by using clear explanations [5]. Dijk et al. found that the performance of gestures related to the speech content helps humans remember the verbs corresponding to the gestures [23].

Gestures are the one of the effective methods of explaining scenes, and many works conduct on this topic. There are several approaches generating gestures to explaining scenes ([12, 19]). Nihei et al. prepared seven types of gestures that express shapes of objects and developed a system that decides the appropriate gesture to represent the shape of an object in an input image [19]. Kadono et al. proposed the system which receives texts as input, checks the pre-prepared dictionary of gestures for each noun obtained from morphological analysis, and performs one of the three pre-prepared gestures if the noun is defined in the dictionary [12].

2.2 Texts Explaining Scenes

Text is also the one of the effective methods of explaining scenes [1, 21, 27, 28]. Image captioning is the task of generating text that describes the content of an image. Selvaraju et al. proposed Gradient-weighted Class Activation Mapping (Grad-CAM) that can grasp each neuron's importance for the decision of interest [21]. Video captioning is the task of generating text that describes the content of a video. Previous research on video captioning uses LSTM or transformer [7, 10, 15, 26]. Yan et al. proposed the STAT video caption framework, which uses the spatial-temporal attention mechanism

(STAT) to exploit the temporal and spatial structure of video [26]. Man et al. proposed a scenario-aware recurrent transformer (SART) that uses a recurrent transformer and includes scenario understanding module [15]. Focusing on the shared memory that a robot and a user acquire during the same experience, Matsumoto et al. developed a computational model of memory recall of visited places, and a robot which responds to a user using the model [17].

2.3 Problem Setting

Previous works do not consider how to select information to convey what it sees to someone who did not see. Humans inherently select information and the representations should be different depending on the person or the situation. In this study, we performed a case study to investigate humans' expressions that conveys what they saw to someone who didn't see it. And, based on the findings given by the case study, we designed the criteria (see Section 4.3) of selecting information from the input and created a system that generates gestures and speech from the information selected according to the criteria.

3 CASE STUDY

3.1 Overview

People commonly communicate something they have seen to people who didn't see. Hence, we aim to build a system that enables a robot to communicate a scenario that it saw to people who did not see that scenario. For a system to generate gestures and speech based on an input video, it must have a function to decide what to talk about in the video. We thus performed a case study to investigate the expressions that Japanese people use to describe video, with the goal of finding classification for judging what they talk about.

3.2 Procedure

We asked experimental participants to watch videos and then explain in one sentence how they would describe each video to someone who had never seen it. The case study was conducted through a Japanese crowdsourcing platform with 30 participants. (20 male, 7 female and 3 undisclosed; age 30–55 years old), who each watched 10 videos. Each video was 10 seconds long and was randomly selected from the video description dataset VATEX [25]. VATEX uses video from Kinetics-600 [6] validation and holdout test sets, and each video has 10 English and 10 Chinese captions. The videos focus on human behavior and include scenes such as playing musical instruments and shaking hands. Expressions for the same thing may change depending on the culture and language. Therefore, translations of these captions could not be used, and descriptions were collected in Japanese.

3.3 Results and Analysis

We collected 300 descriptions in total; Table 1 gives examples from two videos. We analyzed the descriptions to investigate the tendencies in how the participants described the videos. First, we divided the collected descriptions into clauses and extracted the subjects and predicates. Next, we classified the clauses into two groups: those with a relation to the subject and predicate (main idea), and those with no relation. We judged these relations by whether or

Table 1: Examples of collected descriptions


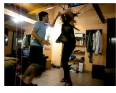
Video	Descriptions (translated from Japanese)
 https://youtu.be/i2F6HrZKo34 [16]	<p>A man kicks another man who is holding crutches with both arms while smoking a cigarette in his mouth.</p> <p>A man standing on crutches is suddenly kicked by another man.</p>
 https://youtu.be/qcP1mb1ljjl [9]	<p>A couple are dancing salsa in a room.</p> <p>A man and a woman are dancing passionately in a room somewhere.</p>

Table 2: Definitions for determining what to talk about

Object to talk about	Definition
Event	An object whose motion has changed from its initial state.
Foreground	An element that is the main idea of the description: either an Event, or something that is not an Event but is moving.
Background	An element that is not the main idea of the description and gives the entire scene.

not the subject and predicate were affected when the clause was deleted. Finally, we examined the tendencies of the subjects and predicates in the descriptions of each video.

As a result, we found three tendencies. First, in about 89% of the descriptions for videos with behavior changes from the initial states, the main ideas in the descriptions involved changes, such as “things fell” or “things stopped.” Second, when there was no change of behavior, the main ideas of about 93% of the descriptions tended to involve motions, such as “dancing” or “moving.” Third, about 32% of the descriptions express elements that had little to do with the main ideas in descriptions. These references to extraneous elements were seemingly intended to make it easier for the person receiving the explanation to picture the scene. From this analysis, we developed the definitions listed in Table 2 and used them to guide our system in determining what to talk about in a video and generating utterances accordingly. We defined Event as an element whose state has changed from its initial state. We defined Foreground as an element which is the main idea of the description. It is an Event, or something that is not an Event but is moving. We defined Background as an element which is not the main idea of the description, but conveys an image of the entire scene. Foreground and Background are referred in combination.

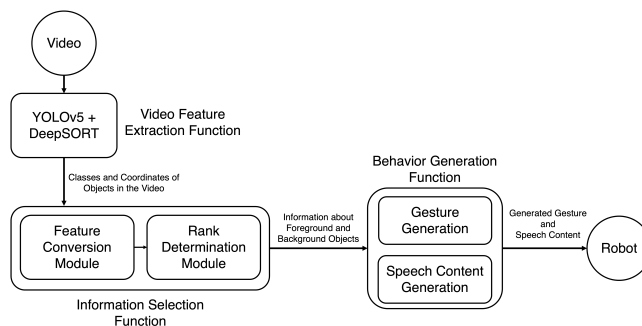


Figure 2: System configuration.

4 VISTURE: SYSTEM FOR SPEECH AND GESTURE GENERATION BASED ON VIDEO

4.1 System Structure

In this paper, we build VISTURE (Video-based Speech and gesture), a gesture and speech generation system for robots that is based on video input. From an input video, VISTURE generates multiple candidate objects to talk about, ranks the candidates, and generates a gesture and utterance that express the chosen object’s motion. VISTURE uses the definitions in Table 2, from the results of our case study, to determine the object to talk about and to generate speech. The system configuration is shown in Figure 2. VISTURE is composed of four functions: a video feature extraction function, an information selection function, and a behavior generation function. When a video is inputted, VISTURE outputs the robot’s gesture and speech.

4.2 Video Feature Extraction Function

This function gets the classes of all the objects in the input video and their coordinates in each frame. For this paper, we used Yolov5 + Deep Sort with PyTorch [4] for object detection and tracking. The classes of the detected objects were based on Microsoft COCO dataset [14].

4.3 Information Selection Function

This function determines the object to talk about by using the definitions in Table 2. Specifically, VISTURE determines the ranking of Foreground and Background candidates to talk about. First, the information selection function receives a set of the detected objects, $o_i, i = 0, \dots, N$, from the video feature extraction function as input. Then, it uses the coordinates of the detected objects to calculate v_{ij} how well they match each criterion $c_j, j = fell, \dots, movingfast, \dots, movingslow, \dots, beinglarge$ via a function $CALCULATE_MATCHING_c_j$:

$$v_{ij} = CALCULATE_MATCHING_c_j(o_i) \quad (1)$$

The criteria are listed in Table 3. Based on the results in Section 3, we devised these criteria that could be computed from the coordinates of the recognized objects. For Event, there may be no applicable object, in which case Event is not taken into account. For each criterion, the function normalizes the degree to which each object fits the criterion, with a value of 1 for the object that fits

Table 3: Criteria for the object to talk about

Classification		Criterion
Foreground	Event	- fell - stopped - started
	Object State	- moving fast - moving long distance - moving from outside the frame to inside
Background		- moving slow - having many objects of the same class - being large

the criterion the best and a value of 0 for the object that fits it the worst. Then, for each criterion, the difference, denoted as $score_j$, between the maximum normalized value m_j and the next highest normalized value n_j is calculated:

$$score_j = m_j - n_j \quad (2)$$

The object that best fits each criterion is chosen as a candidate to talk about. For example, for the criterion “moving long distance,” the information selection function calculates each detected object’s moving distance in relation to its appearance time by $CALCULATE_MATCHING_cmovingdistance(o_i)$. It then normalizes the moving distance and calculates $score_{movinglongdistance}$. The object with the longest moving distance is mentioned when the expression is generated based on this criterion.

As described in Section 3, objects that are Foreground candidates tend to be objects that fit the Event definition. Accordingly, if an object fits the Event definition, it is given priority as a Foreground candidate, regardless of the $score$ values obtained for the non-Event (Object State) criteria.

As noted in Table 2, the Background is not the main point of the description. Therefore, to create combinations of the Foreground and Background candidates from each criterion, after selecting a Foreground candidate, a Background candidate is selected as an object that is recognized as belonging to a different class than the Foreground object. Then, Foreground and Background combinations are created. VISTURE determines that they are suitable for output in order of the sum of their $score$ values:

$$score_{total} = score_{Foreground} + score_{Background} \quad (3)$$

4.4 Behavior Generation Function

4.4.1 Gesture Generation. From the coordinates of the chosen object to talk about, VISTURE generates a gesture for which the robot’s hand position when viewed from the front corresponds to the object’s coordinates. VISTURE assumes that the robot has seen the input video, and that the robot’s hand positions are horizontally flipped from the object’s coordinates. In addition, it adjusts the length of the gesture to match the length of the generated speech, so that the gesture is not much longer or shorter. To match gesture timing with that of the speech, if a gesture’s starting point is far

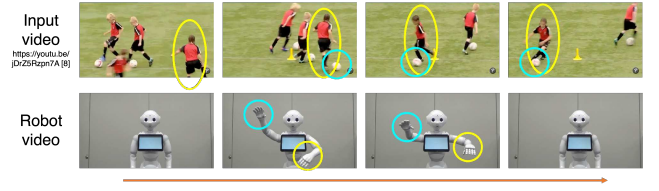


Figure 3: Example of an input video [8] and the robot’s generated gesture. The yellow circle represents the Foreground and the light blue circle represents the Background. The robot uses its right arm to perform a Background gesture and its left arm to perform a Foreground gesture. The class of the Background object is “sports ball,” and its criterion is “moving slow.” The class of the Foreground object is “people,” and its criterion is “stopped.” The objects’ coordinates and the robot’s hand position are reversed because VISTURE assumes a situation in which the robot conveys what it saw and generates the gestures based on the robot’s viewpoint.

from the hand position in an upright state, which is the robot’s default posture, the robot advances to the start of the gesture.

The robot uses an arm to perform a Foreground gesture. When VISTURE uses the “moving slow” and “being large” criteria in Table 3 as the Background classification, the robot uses the opposite arm to describe the Background object, depending on whether the object’s coordinates are closer to the left or right side of the frame, and the hand position corresponds to the object’s coordinates. Then, the robot uses the other arm to describe the Foreground. When “having many objects of the same class” is the Background criterion, the robot uses a metaphoric gesture with arms spread (prepared in advance) as the Background gesture to express a large number of objects, and it performs another gesture to express the Foreground. Metaphoric gesture is the one of the classifications of gestures by McNeill, and represent abstract concepts [18]. We determined this gesture based on McNeill’s classification. When “moving fast” is the Foreground criterion, VISTURE makes the robot move faster than when using the other criteria. In addition, for the “moving long distance” and “moving from outside the frame to inside” criteria, it makes the working range of the robot’s arm larger than for the other criteria. We used SciPy [24] to calculate the angle of the arm joint from the position of the robot’s hand. If we used an object’s coordinates in all frames to generate a gesture, the gesture would be staggered; thus, we average the coordinates every 0.4 seconds. Figure 3 gives an example of an input video and the robot’s generated gesture.

4.4.2 Speech Content Generation. VISTURE generates utterances based on the criteria used to determine the object to talk about and the class of the object. For the class names of the Background and Foreground objects to be talked about, VISTURE combines each one with a sentence that is randomly selected from candidates for each criterion that was used in selecting that object. The generated speech refers to the Background and Foreground. For example, suppose that the class of the Background object is “car,” and that the criterion is “moving slow.” Suppose further that class of the Foreground object is “person,” and that the criterion is “stopped.”

In this case, VISTURE generates a sentence such as “A person stopped, while there was a car in the background.” Specifically, for the Background object, “There was” would be randomly selected from the candidates for the criterion “moving slow,” and “a car” is the class name of the object talked about; thus the resulting sentence is “There was a car.” As for the Foreground object, “is stopped” would be randomly selected from the candidates for the criterion “stopped,” and “a person” is the class name, resulting in the sentence “A person stopped.” Finally, by connecting these two sentences in the order of Background and Foreground, VISTURE generates the sentence “A person stopped, while there was a car in the background.”

4.5 Robot

For the robot, we used Pepper, which was developed by SoftBank Robotics [22]. It has two degree of freedoms in its head, six in its arm, two in its waist, and one in its knee. Pepper is 121 cm tall and weights 29 kg.

5 EXPERIMENTS

5.1 Conditions

We conducted an experiment with an online questionnaire survey to investigate the effects of gesture expression with VISTURE and the findings from our case study. We compared the following three conditions in an experiment having a within-subjects design.

Proposed

The robot performed both the gesture and utterance with the highest ranking among the gestures and utterances generated by VISTURE.

Speech Only

The robot performed the same speech as under the Proposed condition but without a gesture.

Baseline

The robot performed a gesture and utterance without referring to the Background. It talked about the same object as under the Proposed and Speech Only conditions, and it used Object State as the selection criterion for the Foreground object. We used this condition to measure the effect of Background and Event mentions, as found in the case study.

5.2 Procedure

We explained to the participants that the robot would express a scene that it had actually seen by focusing on the motion in that scene. Participants watched three videos generated from an input video for the three experimental conditions, as well as the input video. The participants first watched a video of the robot’s performance. They then watched the input video, which was assumed to be the actual scene viewed by the robot, and they answered a questionnaire. The sequence of steps from watching the video of the robot to answering the questionnaire was repeated for each video under each of the three experimental conditions. We counterbalanced the order of the presented videos.

For the input videos, we randomly selected four from the video description dataset VATEX. As described in Section 3, VISTURE uses objects recognized in different classes as the Foreground and

Table 4: Questionnaire 1 (translated from Japanese)

Measure	Item	Questionnaire content
Gesture	Q1	The robot’s gesture was appropriate for the input video.
	Q2	The robot’s gesture was appropriate for the speech content.
	Q3	The robot’s gesture was natural.
Speech	Q4	The robot’s speech was appropriate for the input video.
	Q5	The robot’s speech was appropriate for the gesture.
	Q6	The robot’s speech was natural.
Overall representation	Q7	The robot’s explanation was appropriate for the input video.
	Q8	I understood what the robot was trying to tell me.
	Q9	The robot’s explanation was sufficient for the input video.

Background. Accordingly, we used videos in which more than two objects appeared and at least two or more classes of objects were recognized. From these four input videos, VISTURE generated expressions that mentioned the Foreground object, which was determined according to the Event criteria in all the videos.

5.3 Measurements

We prepared Questionnaire 1, listed in Table 4, to investigate the participants’ impressions of the robot’s gestures and utterances. The items on Questionnaire 1 were evaluated on a 7-point Likert scale, where 1 was the most negative response and 7 was the most positive. We also used the Godspeed questionnaire [2], which is a scale to evaluate the perceived impressions of robots from five perspectives: anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety. We used all of these perspectives except the perceived safety. The Godspeed questionnaire items were evaluated on a 5-point Likert scale, where 1 was the most negative response and 5 was the most positive.

5.4 Expected Results

We expected that the responses for anthropomorphism and likeability would be higher under the Proposed and Baseline conditions than under the Speech Only condition, because Salem et al. found that people perceive a robot speaking with gestures as more human-like and likeability [20]. Accordingly, we made the following prediction:

Prediction 1: As compared to the Speech Only condition, the robot’s performance under the Proposed and Baseline conditions will be rated higher on the Godspeed questionnaire.

We also expected that the robot’s likeability would be increased by taking the Background and Event into account, i.e., under the Proposed condition. In the case study, the majority of participants mentioned an Event as Foreground when an Event occurred. This suggests that inclusion of the Background and Event can more

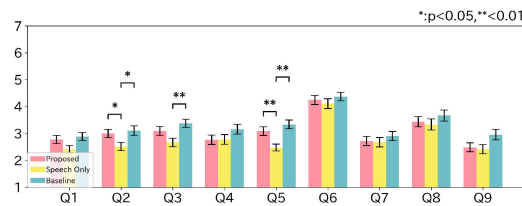


Figure 4: Results of Questionnaire 1 for all videos. The error bars indicate the standard errors.

Table 5: Results of Questionnaire 1 for all videos.

Item	F	p	η^2
Q1	$F(1.819, 141.890) = 3.021$	0.057	0.04
Q2	$F(1.859, 145.009) = 4.167$	0.020	0.05
Q3	$F(2, 156) = 5.630$	0.004	0.07
Q4	$F(1.416, 110.417) = 2.214$	0.130	0.03
Q5	$F(2, 156) = 9.735$	0.001	0.11
Q6	$F(1.763, 137.513) = 1.338$	0.265	0.02
Q7	$F(1.526, 118.995) = 0.722$	0.453	0.01
Q8	$F(1.368, 106.728) = 1.134$	0.308	0.01
Q9	$F(1.235, 96.359) = 3.849$	0.044	0.05

strongly express the characteristics of a video, and we thus made the following prediction:

Prediction 2: As compared to the Baseline condition, the robot’s performance under the Proposed condition will be rated higher on both Questionnaire 1 and the Godspeed questionnaire.

5.5 Participants

Through a Japanese crowdsourcing service, we recruited 20 participants for each input video, giving a total of 80 participants for four input videos (51 male, 20 female, 9 undisclosed; age 20–65 years old). We removed one participant from the analysis as an outlier because the participant’s response time for the questionnaire was 308 s, whereas the median response time was 664 s.

6 RESULTS

Figure 4 shows the results of Questionnaire 1. We conducted a one-way repeated-measures ANOVA on the results. What kind of video was used is not treated as an independent variable. It revealed a significant effect for Q2, Q3, Q5, and Q9 between the conditions. In contrast, there was no significant effect for Q1, Q4, Q6, Q7, or Q8. Table 5 shows details.

We also conducted multiple comparisons using the Bonferroni method, which revealed significant differences for Q2 between the Proposed and Speech Only conditions ($p = 0.035$), and between the Speech Only and Baseline conditions ($p = 0.038$). For Q3, we found a significant difference between the Speech Only and Baseline conditions ($p = 0.003$). For Q5, there were significant differences between the Proposed and Speech Only conditions ($p = 0.005$), and between the Speech Only and Baseline conditions ($p < 0.001$). Lastly, there were no significant differences for Q9.

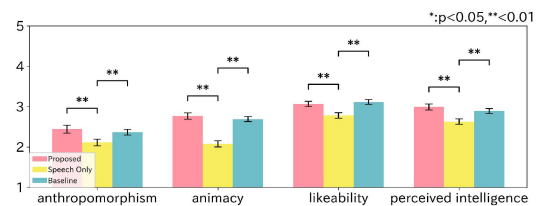


Figure 5: Results of the Godspeed questionnaire for all videos. The error bars indicate the standard errors.

Table 6: Results of the Godspeed questionnaire for all videos.

Item	F	p	η^2
anthropomorphism	$F(2, 156) = 7.269$	0.001	0.09
animacy	$F(2, 156) = 40.233$	0.001	0.34
likeability	$F(2, 156) = 15.598$	0.001	0.17
perceived intelligence	$F(1.765, 137.640) = 10.409$	0.001	0.12

Next, Figure 5 shows the results of the Godspeed questionnaire. We found a significant effect between the conditions for all four tested perspectives. Table 6 shows details.

For the Godspeed questionnaire, we again conducted multiple comparisons using the Bonferroni method. We found significant differences for anthropomorphism between the Proposed and Speech Only conditions ($p = 0.002$), and between the Speech Only and Baseline conditions ($p = 0.008$). For animacy, likeability, perceived intelligence, there were significant differences between the Proposed and Speech Only conditions ($p < 0.001$), and between the Speech Only and Baseline conditions ($p < 0.001$).

7 DISCUSSION

7.1 Implications

For all items on the Godspeed questionnaire, we found significant differences between the Proposed and Speech Only conditions, and between the Speech Only and Baseline conditions. Because the robot’s performance under the Proposed and Baseline conditions was rated higher than under the Speech Only condition, Prediction 1 was validated. VISTURE generates gestures that express motion, and we found that the participants’ impressions of the robot were enhanced when a gesture expressing motion was performed along with speech. In other words, it is worthwhile for a robot to perform gestures that express motion.

On the other hand, we observed no trend for robot expressions that mimicked human expressions of the Background and Event. In other words, we could not validate Prediction 2. However, by examining the trends for the individual videos, we found cases in which expressions that included the Background and Event were evaluated better or worse. We describe these findings in the next section.

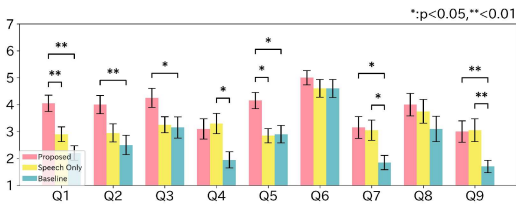


Figure 6: Results of Questionnaire 1 for Video 1. The error bars indicate the standard errors.

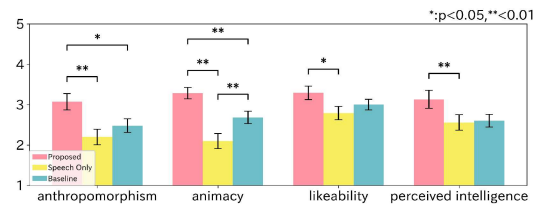


Figure 7: Results of the Godspeed questionnaire for Video 1. The error bars indicate the standard errors.

7.2 Effects of Events and Background Descriptions

Here, we discuss when a representation taking the Event and Background into account works well, when it does not, and what kind of representation should be used. The input video affected whether the robot’s performance was evaluated more highly under the Proposed condition or the Baseline condition; hence, we focus on this difference.

For all of the input videos used in the experiment, VISTURE generated representations in which the Foreground object was obtained from an Event. Therefore, the difference between the Foreground representations under the Proposed and Baseline conditions was whether or not the target object to talk about was obtained from an Event, and whether or not the generated expression took the Background into account.

For Video 1 of a man throwing a ball, the robot’s performance was evaluated more positively under the Proposed condition than under the Baseline condition. By conducting a one-way repeated-measures ANOVA and multiple comparisons using the Bonferroni method, we found significant differences between the Proposed and Baseline conditions. Figure 6 shows the results of the multiple comparisons test for Questionnaire 1. The test identified the following results: for Q1, Proposed > Speech Only ($p = 0.006$) and Proposed > Baseline ($p < 0.001$); for Q2, Proposed > Baseline ($p = 0.006$); for Q3, Proposed > Baseline ($p = 0.044$); for Q4, Speech Only > Baseline ($p = 0.035$); for Q5, Proposed > Speech Only ($p = 0.017$) and Proposed > Baseline ($p = 0.011$); for Q7, Speech Only > Baseline ($p = 0.028$) and Proposed > Baseline ($p = 0.027$); and for Q9, Speech Only > Baseline ($p = 0.004$) and Proposed > Baseline ($p = 0.001$). Overall, these results show that the robot’s performance for Video 1 was better evaluated under the Proposed condition than under the Baseline condition.

Next, Figure 7 shows the results of the multiple comparisons test for the Godspeed questionnaire. The test identified the following results: for anthropomorphism, Proposed > Speech Only ($p = 0.002$) and Proposed > Baseline ($p = 0.010$); for animacy, Proposed > Speech Only ($p < 0.001$), Speech Only > Baseline ($p = 0.001$), and Proposed > Baseline ($p < 0.001$); for likeability, Proposed > Speech Only ($p = 0.039$); and for perceived intelligence, Proposed > Speech Only ($p = 0.009$). Overall, these results further indicate that the performance for Video 1 was better evaluated under the Proposed condition than under the Baseline condition.

To a human observer, Video 1 appears to show an Event occurring during the video. The intent of mentioning the Background is

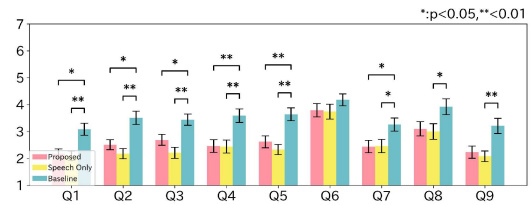


Figure 8: Results of Questionnaire 1 for Videos 2 and 3. The error bars indicate the standard errors.

to make the explanation more detailed, and it is thus better to do so than not to do so. As a result, because there was no discrepancy between Video 1 and the representation generated by VISTURE, the robot’s performance was better evaluated under the Proposed condition, which provided a more detailed explanation than under the Baseline condition.

In contrast, for Video 2 of boys practicing dribbling and Video 3 of boys riding a scooter, the robot’s performance was evaluated more positively under the Baseline condition than under the Proposed condition. We again conducted a one-way repeated-measures ANOVA and multiple comparisons using the Bonferroni method, and we found significant differences between the Proposed and Baseline conditions. Figure 8 shows the results of the multiple comparisons test for Questionnaire 1. The test identified the following results: for Q1, Baseline > Speech Only ($p = 0.004$) and Baseline > Proposed ($p = 0.011$); for Q2, Baseline > Speech Only ($p = 0.001$) and Baseline > Proposed ($p = 0.010$); for Q3, Baseline > Speech Only ($p < 0.001$) and Baseline > Proposed ($p = 0.049$); for Q4, Baseline > Speech Only ($p = 0.001$) and Baseline > Proposed ($p = 0.003$); for Q5, Baseline > Speech Only ($p < 0.001$) and Baseline > Proposed ($p = 0.002$); for Q7, Baseline > Speech Only ($p = 0.031$) and Baseline > Proposed ($p = 0.036$); for Q8, Baseline > Speech Only ($p = 0.023$); and for Q9, Baseline > Speech Only ($p = 0.007$). Overall, the robot’s performance for Videos 2 and 3 was better evaluated under the Baseline condition than under the Proposed condition.

Next, Figure 9 shows the results of the multiple comparisons test for the Godspeed Questionnaire. The test identified the following results: for animacy, Proposed > Speech Only ($p = 0.001$) and Baseline > Speech Only ($p < 0.001$); for likeability, Proposed > Speech Only ($p = 0.006$) and Baseline > Speech Only ($p < 0.001$); and for perceived intelligence, Baseline > Speech Only ($p < 0.001$) and

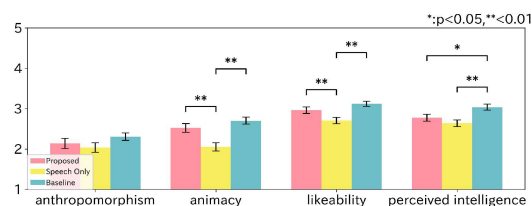


Figure 9: Results of the Godspeed questionnaire for Videos 2 and 3. The error bars indicate the standard errors.

Baseline > Proposed ($p = 0.013$). Overall, these results further confirm that the performance for Videos 2 and 3 was better evaluated under the Baseline condition than under the Proposed condition.

To a human observer, in contrast to Video 1, Videos 2 and 3 do not appear to show an Event occurring during the video. It seems that there was a representation mismatch under the Proposed condition, whereas the mismatch did not occur under the Baseline condition because the Event was not considered. The main cause of the representation mismatch seems to be that VISTURE generates representations by using only coordinate changes. Comparing the VISTURE output results with the input Videos 2 and 3, we can observe that there were objects whose coordinates changed very little in the videos. As a result, the “stopped” criterion for an Event was used to determine the target object to talk about. However, the objects were still moving in the actual input videos, though the coordinate changes were small. For example, Video 2 shows boys kicking a soccer ball for dribbling practice and trying to go around a colored cone. Because the boys’ coordinates obtained by the video feature extraction function changed very little when they tried to go around the cone, the information selection function decided that the “stopped” criterion for an Event was satisfied. Thus, we found that it is not possible to detect all human actions from coordinate changes alone.

The robot’s representation taking an Event into account did not necessarily improve its evaluation by the participants. Expressions that take an Event into account may be effective for objects such as cars, whose actions are directly related to coordinate changes. However, humans may perform actions without changing coordinates, in which case taking an Event into account may result in representation mismatches. For the case when the target Foreground object is a human, we should consider either generating a representation that does not take an Event into account or introducing an action classification. This knowledge of meta-level rule can also bring useful insights to architectural design when creating models in deep learning.

7.3 Limitations

We note here that our study has some limitations. The robot’s gesture generation controls only the hand’s position and the range of motion of the robot’s arms is limited to two dimensions. It is because its gestures are generated from the coordinates of objects in a video, but gestures are normally three-dimensional. It would be interesting to conduct experiments that focus more on the effects of gestures, such as comparing speech-related gestures with random gestures. Regarding the limitations of the experimental design, the

results may not be generalizable across cultures, as expressions vary among different languages and cultures. In addition, the experiment was conducted with only one type of robot. Because Pepper has a human-like appearance, different trends may emerge when using a robot with a simpler appearance. Although the viewpoints of the input videos should be an important factor because a robot is intended to describe a view that it has seen, we did not account for it in this study. Our experimental design limited the possible types of communication. In the future, we aim to investigate more interactive settings by using real-world information.

8 CONCLUSION

We have proposed VISTURE, a system that generates robot gestures and speech by using video as input, assuming a situation in which a robot conveys what it saw with a camera to a person who was absent. We have performed a case study to investigate the expressions that Japanese people use to describe video scenes and found classification: Foreground and Background. VISTURE generates representations based on changes in the coordinates of objects in a video, according to the findings of a case study. We experimentally evaluated the quality of the gestures and speech generated by VISTURE and the people’s impression of a robot performing those gestures and speech. We found that the robot was perceived as more likeable and capable when it performed gestures. However, some representations that took an Event and the Background in a video into account worked well, whereas others did not. This was because coordinate changes in a video could not always be used to detect human actions, in which case the generated representation mentioning an Event did not match the person performing an action.

ACKNOWLEDGMENTS

This work was supported in part by JST, CREST Grant Number JPMJCR19A1, Japan and JSPS KAKENHI Grant Number JP20K19897.

REFERENCES

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6077–6086. <https://doi.org/10.1109/CVPR.2018.00636>
- [2] Christoph Bartneck, Elizabeth Croft, and Dana Kulic. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81. <https://doi.org/10.1007/s12369-008-0001-3>
- [3] Paul Bremner and Ute Leonards. 2016. Iconic Gestures for Robot Avatars, Recognition and Integration with Speech. *Frontiers in Psychology* 7 (2016). <https://doi.org/10.3389/fpsyg.2016.00183>
- [4] Mikel Broström. 2020. Real-time multi-object tracker using YOLOv5 and deep sort. https://github.com/mikel-brostrom/Yolov5_DeepSort_Pytorch.
- [5] John-John Cabibihan, Wing-Chee So, Sujin Saj, and Zhengchen Zhang. 2012. Telerobotic Pointing Gestures Shape Human Spatial Cognition. *International Journal of Social Robotics* 4 (April 2012), 263–272. <https://doi.org/10.1007/s12369-012-0148-9>
- [6] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A Short Note about Kinetics-600. <https://doi.org/10.48550/ARXIV.1808.01340>
- [7] Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Trevor Darrell. 2017. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2017), 677–691. <https://doi.org/10.1109/TPAMI.2016.2599174>
- [8] Drills. 2016. Soccer Coaching Attacking Drill: Warm Up | circular saw. <https://youtu.be/jDrZ5Rzpn7A> Accessed: Jun 18, 2022.

- [9] Martin Fernandez. 2013. Dami y Barby Bailando Salsa. <https://youtu.be/qcP1mb1ljjI> Accessed: Jun 18, 2022.
- [10] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. 2017. Video Captioning With Attention-Based LSTM and Semantic Consistency. *IEEE Transactions on Multimedia* 19, 9 (2017), 2045–2055. <https://doi.org/10.1109/TMM.2017.2729019>
- [11] Michita Imai, Tetsuo Ono, and Hiroshi Ishiguro. 2001. Physical relation and expression: joint attention for human-robot interaction. In *Proceedings 10th IEEE International Workshop on Robot and Human Interactive Communication. ROMAN 2001 (Cat. No.01TH8591)*. 512–517. <https://doi.org/10.1109/ROMAN.2001.981955>
- [12] Yuki Kadono, Yutaka Takase, and Yukiko I. Nakano. 2016. Generating Iconic Gestures Based on Graphic Data Analysis and Clustering. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction (Christchurch, New Zealand) (HRI '16)*. IEEE Press, 447–448.
- [13] Mitsuhiro Kimoto, Takamasa Iio, Masahiro Shiomi, and Katsunori Shimohara. 2021. Coordinating Entrainment Phenomena: Robot Conversation Strategy for Object Recognition. *Applied Sciences* 11, 5 (2021). <https://doi.org/10.3390/app11052358>
- [14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.
- [15] Xin Man, Deqiang Ouyang, Xiangpeng Li, Jingkuan Song, and Jie Shao. 2022. Scenario-Aware Recurrent Transformer for Goal-Directed Video Captioning. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 4, Article 104 (mar 2022), 17 pages. <https://doi.org/10.1145/3503927>
- [16] Nicholas Masters. 2011. Guy on crutches gets dropped kicked. <https://youtu.be/i2F6HrZKo34> Accessed: Jun 18, 2022.
- [17] Takahiro Matsumoto, Satoru Satake, Takayuki Kanda, Michita Imai, and Norihiro Hagita. 2012. Do you remember that shop? – Computational model of spatial memory for shopping companion robots. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 447–454.
- [18] David McNeill. 1992. *Hand and mind : what gestures reveal about thought*. University of Chicago Press.
- [19] Fumio Nihei, Yukiko Nakano, Ryuichiro Higashinaka, and Ryo Ishii. 2019. Determining Iconic Gesture Forms Based on Entity Image Representation. In *2019 International Conference on Multimodal Interaction (Suzhou, China) (ICMI '19)*. Association for Computing Machinery, New York, NY, USA, 419–425. <https://doi.org/10.1145/3340555.3353736>
- [20] Maha Salem, Friederike Eyszel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability. *International Journal of Social Robotics* 5 (2013), 313–323.
- [21] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626. <https://doi.org/10.1109/ICCV.2017.74>
- [22] SoftBank. 2019. Robot SoftBank. <https://www.softbank.jp/en/robot/> Accessed: Jun 18, 2022.
- [23] Elisabeth T Van Dijk, Elena Torta, and Raymond H Cuijpers. 2013. Effects of eye contact and iconic gestures on message retention in human-robot interaction. *International Journal of Social Robotics* 5, 4 (2013), 491–501. <https://doi.org/10.1007/s12369-013-0214-y>
- [24] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, Ilhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. <https://doi.org/10.1038/s41592-019-0686-2>
- [25] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *The IEEE International Conference on Computer Vision (ICCV)*.
- [26] Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai. 2020. STAT: Spatial-Temporal Attention Mechanism for Video Captioning. *IEEE Transactions on Multimedia* 22, 1 (2020), 229–241. <https://doi.org/10.1109/TMM.2019.2924576>
- [27] Xu Yan, Zhengcong Fei, Zekang Li, Shuhui Wang, Qingming Huang, and Qi Tian. 2021. *Semi-Autoregressive Image Captioning*. Association for Computing Machinery, New York, NY, USA, 2708–2716. <https://doi.org/10.1145/3474085.3475179>
- [28] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning with Semantic Attention. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4651–4659. <https://doi.org/10.1109/CVPR.2016.503>