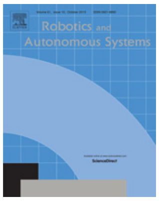
Robotics and Autonomous Systems 62 (2014) 721–736

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot



Multimodal integration learning of robot behavior using deep neural networks

Kuniaki Noda*, Hiroaki Arie, Yuki Suga, Tetsuya Ogata

Department of Intermedia Art and Science, Graduate School of Fundamental Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo, 169-8555, Japan

HIGHLIGHTS

- Novel computational framework for sensory-motor integration learning.
- Cross-modal memory retrieval utilizing a deep autoencoder.
- Noise-robust behavior recognition utilizing acquired multimodal features.
- Multimodal causality acquisition and sensory-motor prediction.

ARTICLE INFO

Article history:
Received 8 October 2013
Received in revised form
20 January 2014
Accepted 15 March 2014
Available online 24 March 2014

Keywords:
Object manipulation
Multimodal integration
Cross-modal memory retrieval
Deep learning

ABSTRACT

For humans to accurately understand the world around them, multimodal integration is essential because it enhances perceptual precision and reduces ambiguity. Computational models replicating such human ability may contribute to the practical use of robots in daily human living environments; however, primarily because of scalability problems that conventional machine learning algorithms suffer from, sensory-motor information processing in robotic applications has typically been achieved via modal-dependent processes. In this paper, we propose a novel computational framework enabling the integration of sensory-motor time-series data and the self-organization of multimodal fused representations based on a deep learning approach. To evaluate our proposed model, we conducted two behavior-learning experiments utilizing a humanoid robot; the experiments consisted of object manipulation and bell-ringing tasks. From our experimental results, we show that large amounts of sensory-motor information, including raw RGB images, sound spectrums, and joint angles, are directly fused to generate higher-level multimodal representations. Further, we demonstrated that our proposed framework realizes the following three functions: (1) cross-modal memory retrieval utilizing the information complementation capability of the deep autoencoder; (2) noise-robust behavior recognition utilizing the generalization capability of multimodal features; and (3) multimodal causality acquisition and sensory-motor prediction based on the acquired causality.

1. Introduction

Humans are known to perceive the external environment, including their own body, by utilizing multiple sources of sensory information, such as vision, audition, and proprioception. To this end, multimodal integration contributes to forming constant, coherent, and robust perceptions by reducing ambiguities regarding sensory environment. Cognitive science research has revealed that by combining sensory information, humans achieve enhanced perceptual clarity and reduced ambiguity regarding their environment [1,2]. Further, action–effect causality perception is known to have a close relationship with the sense of agency [3], and thus cross-modal grouping plays an important role for sensation [4]. Hence, we believe that replicating human multimodal integration learning as a computational model is essential toward realizing sophisticated cognitive functions of robot intelligence, as well as toward fundamentally understanding human intelligence.

Unfortunately, multimodal integration has long been a challenging problem in robotics [5,6]. Although there is relevant research reported in the literature [7–9], several issues still remain unsolved. First, multimodal sensory-motor integration has

* Corresponding author. Tel.: +81 3 5286 2742.
E-mail addresses: kuniaki.noda@akane.waseda.jp (K. Noda), arie@aoni.waseda.jp (H. Arie), ysuga@ysuga.net (Y. Suga), ogata@waseda.jp (T. Ogata).

http://dx.doi.org/10.1016/j.robot.2014.03.003

typically been applied only to a singular problem, such as self-organizing one's spatial representation [7,8]; further functions have not been intensively studied, including such functions as the cross-modal complementation of information deficiencies or the application of cross-modal memory retrieval for behavior generation problems. Second, discussion in the literature regarding how multimodal information should be fused together to realize stable environmental recognition has not reached a comprehensive consensus. Thus, a prevailing multimodal information integration framework has not been available. Subsequently, in robotics, sensory inputs acquired from different sources are still typically processed with dedicated feature-extraction mechanisms [10]. Third, multimodal causality modeling as a means to implementing sensory-motor prediction for robotic applications has not been adequately investigated. Several preceding studies have proposed computational models developmentally acquiring action–effect causality toward understanding interaction rules [11,12]; however, most causal models have been represented using a limited number of modalities, often focused on vision and motion.

A scalable learning framework that enables multimodal integration learning by handling large amounts of sensory-motor data with high dimensionality has not yet been realized. In line with the growing demand for perceptual precision in regard to the surrounding environment, recent robots are equipped with state-of-the-art sensory devices, such as high-resolution image sensors, range sensors, multichannel microphones, and so on [13–15]. As a result, remarkable improvements have been achieved in the quantity of available sensory information; however, because of scalability limitations of conventional machine learning algorithms, groundbreaking computational models achieving robust behavior control and environmental recognition by fusing multimodal sensory inputs into a single representation have not yet been proposed.

Regarding computational models addressing large-scale data processing with significant dimensionality [16], deep learning approaches have recently attracted increasing attention in the machine-learning community [17]. For example, deep neural networks (DNNs) have successfully been applied to unsupervised feature learning for single modalities, such as text [18], images [19], or audio [20]. In such studies, various information signals, even with high-dimensional representations, were effectively compressed in a restorable form. Further, brilliant achievements in deep learning technologies have already succeeded in making advanced applications available to the public. For example, competition results from the ImageNet Large Scale Visual Recognition Challenge [21] have led to significant improvements in web image search engines [22]. As another example, unsupervised feature-extraction functions of deep learning technologies have greatly increased the sophistication of a voice recognition engine used for a virtual assistant service [23]. The same approach has also been applied to the learning of fused representations over multiple modalities, resulting in significant improvements in speech recognition performance [24]. Yet another study on multimodal integration learning has succeeded in cross-modal memory retrieval by complementing missing modalities [25]. Most current studies on multimodal integration learning utilize deep networks; however, much work focuses in extracting correlations between static modalities, such as image and text [21]. Thus, few studies have investigated methods not only for multimodal sensor fusion, but also for dynamic sensory-motor coordination problems [26] of robot behavior.

Our interest in this current study is to investigate the fundamental principles reported by the cognitive science studies and put their findings to practical use by constructing a computational model. In particular, we set out to demonstrate the following functions of our proposed model through experimentation on the sensory-motor coordination learning of robot behavior:

- Cross-modal memory retrieval utilizing the information complementation capability of the deep autoencoder.
- Noise-robust behavior recognition utilizing the generalization capability of multimodal features.
- Multimodal causality acquisition and sensory-motor prediction based on the acquired causality.

As a practical computational model, we construct a multimodal temporal sequence learning framework based on a deep learning algorithm [27]. The proposed framework first compresses the dimensionality of the sensory inputs acquired from multiple modalities utilizing the same techniques as previous work on dimensionality compression via a deep autoencoder [27,28]. In combination with a variant of a time-delayed neural network [29] learning approach, we then introduce a novel deep learning framework that integrates sensory-motor sequences and self-organizes higher-level multimodal features. Further, we show that our proposed temporal sequence learning framework can internally generate temporal sequences by partially masking the input data from outside the network and recursively feeding back the previous outputs to the masked inputs nodes, which is made possible by utilizing the characteristics of an autoencoder that models identity mappings between inputs and outputs.

We evaluate our proposed sensory-motor integration learning framework via two behavior learning experiments. First, we train our proposed model with six different object manipulation behaviors of a humanoid robot, generated by direct teaching. Results demonstrate that our proposed method can retrieve temporal sequences over visual and motion modalities and predict future sequences from the past. Further, behavior-dependent unified representations that fuse sensory-motor modalities together are extracted in the temporal sequence feature space. Our behavior recognition experiment, which utilizes the integrated features acquired from the multimodal temporal sequence learning mechanism, demonstrates that the multimodal features significantly improve the robustness and reliability of behavior recognition performance by utilizing joint angle information.

In the second experiment, we extend the multimodal integration learning by incorporating sound signals in addition to the image and joint angles. We designed a bell-ringing task performed by the same robot and trained the proposed model utilizing sensory-motor sequences consist of the three modalities. To this end, a model representing the cross-modal causal dependency is self-organized in the abstracted feature space of our proposed model. Results demonstrated that the cross-modal memory retrieval function of our proposed model succeeds in predicting visual sequences in correlation with the sound and joint angles of bell-ringing behaviors. Further, analyzing image retrieval performance, we found that our proposed method correctly models the causal dependencies among the multimodal information.

In addition to this introductory section, our paper is organized as follows. In Section 2, we briefly review Hessian-free optimization for training deep networks. In Section 3, we describe the general framework of multimodal temporal sequence learning. In Section 4, we present the practical application and effectiveness of our proposed model by the multimodal integration learning of object manipulation behavior utilizing a humanoid robot. Next, in Section 5, we move on to the bell-ringing task, in which we demonstrate how our proposed model acquires the causality between multiple modalities and supports cross-modal information retrieval based on the causal relationship among multiple modalities. In Section 6, we examine and discuss our proposed framework and results in relation to previous work. Finally, we conclude our work in Section 7.

2. Deep neural networks

2.1. Training deep neural networks

Deep neural networks (DNNs) are artificial neural network models with multiple layers of hidden units between inputs and outputs. Hinton et al. first proposed an unsupervised learning algorithm to use greedy layer-wise unsupervised pretraining followed by fine-tuning methods for overcoming high prevalence of unsatisfactory local optima in learning objectives of deep models [27]. Subsequently, Martens proposed a novel approach by introducing a second-order optimization method – Hessian-free optimization – for training deep networks [28]. The proposed method efficiently trained the models by a general optimizer without pretraining. Here, we adopt learning methods proposed by Martens for optimizing multiple autoencoders, for the self-organization of feature vectors, and for temporal sequence learning.

2.2. Hessian-free optimization

The Hessian-free algorithm originates with Newton's method, a well-known numerical optimization technique. A canonical second-order optimization scheme such as Newton's method iteratively updates parameter $\theta \in \mathbb{R}^N$ of an objective function f by computing gradient vector p, and updates θ as $\theta_{n+1} = \theta_n + \alpha p_n$ with learning parameter α. The core idea of Newton's method is to locally approximate f around each θ, up to the second order, by quadratic equation,

$$M_{\theta_n}(\theta) \equiv f(\theta_n) + \nabla f(\theta_n)^T p_n + \frac{1}{2} p_n^T B_{\theta_n} p_n, \tag{1}$$

where B_{θ_n} is a damped Hessian matrix of f at θ_n. As H can become indefinite, the Hessian matrix is re-conditioned to be $B_{\theta_n} = H(\theta_n) + \lambda I$, where $\lambda \geq 0$ is a damping parameter and I is the unit matrix.

Using the standard Newton's method, $M_{\theta_n}(\theta)$ is optimized by computing $N \times N$ matrix B_{θ_n} then solving system $B_{\theta_n} p_n = -\nabla f(\theta_n)^T$. This computation, however, is very expensive for large N, which is a common case even with modestly sized neural networks. To overcome this issue, the variant of Hessian-free optimization developed by Martens utilizes the linear conjugate gradient (CG) algorithm for optimizing quadratic objectives in combination with the use of a positive semidefinite Gauss–Newton curvature matrix in place of the possibly indefinite Hessian matrix. The name "Hessian-free" indicates that the CG does not necessarily require the costly explicit Hessian matrix; instead, the matrix–vector product between the Hessian matrix H or the Gauss–Newton matrix G and gradient vector p is sufficient (for more details on the concrete implementation, see [28,30,31]).

3. Multimodal temporal sequence learning mechanism

In this paper, we propose to apply a deep autoencoder not only for its feature extraction by dimensionality compression but also for its multimodal temporal sequence integration learning. Our main contribution in this study is to demonstrate that our proposed framework serves as a cross-modal memory retriever, as well as a temporal sequence predictor utilizing its powerful generalization capabilities. In the subsections that follows, we first illustrate the basic mechanism of the autoencoder, then explain how the autoencoder is applied for its multimodal temporal sequence learning and further functions.

3.1. Self-organization of feature vectors

High-dimensional raw sensory inputs, such as visual images or sound spectrums, can be converted to low-dimensional feature vectors by multilayer networks with a small central layer (i.e. a feature-extraction network) [27]. To this end, the networks are trained with the goal of reconstructing the input data at the output layer with input–output mappings defined as

$$u_t = f(r_t) \tag{2}$$

$$\hat{r}_t = f^{-1}(u_t), \tag{3}$$

where r_t, u_t, and $\hat{r}_t$ are the vectors representing the raw input data, the corresponding feature, and the reconstructed data, respectively. Functions $f(.)$ and $f^{-1}(.)$ represent the transformation mapping from the input layer to the central hidden layer and the central hidden layer to the output layer of the network, respectively. An autoencoder compresses the dimensionality of inputs by decreasing the number of nodes from the input layer to the central hidden layer. Hence, the number of central hidden layer nodes determines the dimension of the feature vector. In a symmetric fashion, the original input is reconstructed from the feature vector by eventually increasing the number of nodes from the central hidden layer to the output layer.

Regarding dimensionality compression mechanisms, a simple and commonly utilized approach is principal component analysis (PCA); however, Hinton et al. demonstrated that the deep autoencoder outperformed PCA in image reconstruction and compressed feature acquisition [27]. In reference to their work, we utilized the deep autoencoder for our dimensionality compression framework because we prioritized the precision of cross-modal memory retrieval and the sparseness of acquired features to ease the behavior recognition task via a conventional classifier.

3.2. Multimodal learning of temporal sequence using time-delay networks

A time-delay neural network (TDNN) is a method for utilizing a feed-forward neural network for multi-dimensional temporal sequence learning [29]. Motivated by TDNN, we propose a novel computational framework that utilizes a deep autoencoder for temporal sequence learning.

An input to the temporal sequence learning network at a single time step is defined by a time segment of the tuple of joint angle vectors, image feature vectors, and sound feature vectors, formatted as

$$s_t = (a_{\mathbf{t}}, u^i_{\mathbf{t}}, u^s_{\mathbf{t}}) \tag{4}$$

$$\{\mathbf{t} | t - T + 1 \leq \mathbf{t} \leq t\}, \tag{5}$$

where s_t, a_t, u^i_t, and u^s_t are the vectors representing the input to the network, the joint angle, the image feature, and the sound feature, at time t, respectively, and T is the length of the time window. Here, $\mathbf{t}$ represents the previous T steps of the temporal segment from t, and a vector with subscript $\mathbf{t}$ indicates a time series of the vector. The input–output mappings of the temporal sequence learning network are defined as

$$v_t = g(s_t) \tag{6}$$

$$\hat{s}_t = g^{-1}(v_t), \tag{7}$$

where v_t and $\hat{s}_t = (\hat{a}_{\mathbf{t}}, \hat{u}^i_{\mathbf{t}}, \hat{u}^s_{\mathbf{t}})$ are the multimodal feature vector and a segment of the restored multimodal temporal sequence, respectively. Functions $g(.)$ and $g^{-1}(.)$ represent the transformation mapping from the input layer to the central hidden layer and the central hidden layer to the output layer of the network, respectively.

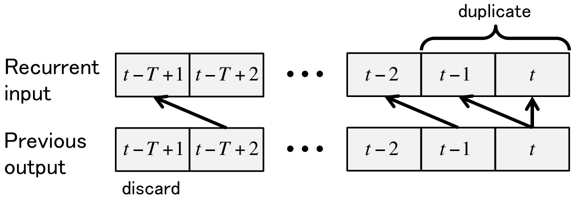


Fig. 1. Buffer shift of the recurrent input.

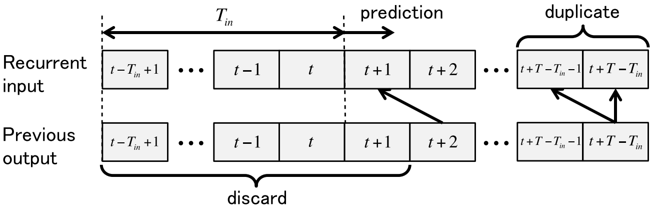


Fig. 2. Buffer shift of the recurrent input for temporal sequence prediction.

One of the merits of applying neural networks for multimodal temporal sequence learning is their generalization capability. Because the network can complement deficiencies in the input data, the temporal sequence learning network can be used in two different ways: (1) to retrieve a temporal sequence from one modal for use in another and (2) to predict a future sequence from the past sequence. Thus, the temporal sequence learning network serves as a cross-modal memory retriever or a temporal sequence predictor by masking the input data from outside the network in either spatial or temporal ways; thus iteratively feeding back the generated outputs to the inputs as substitutions for the masked inputs. The practical implementation of these functions is described in the following subsections.

3.3. *Cross-modal memory retrieval*

Cross-modal memory retrieval is realized by self-generating sequences for a modality inside the network by providing corresponding sequences for the other modalities from outside the network. For the retrieved modality, a recurrent loop from the output nodes to the input nodes is prepared. Hence, in the case of generating an image sequence from motion and sound sequences, input to the network is defined as

$$s_t = (a_{\mathbf{t}}, \hat{u}^i_{\mathbf{t}}, u^s_{\mathbf{t}}). \tag{8}$$

As shown in Fig. 1, the time segment of the recurrent input is generated by shifting the previous output of the network to the direction for one step by (1) discarding the oldest time step output and (2) filling the latest time step with the value of the newest time step acquired from the output.

3.4. *Temporal sequence prediction*

Similarly, the temporal sequence prediction is realized by constructing a recurrent loop from the output layer to the input layer. The difference is that among all the T steps of the time window, only the first T_{in} steps (i.e., the past T_{in} shifts to the present time step t) of both modalities are filled with the input data; the rest (i.e., the future $T - T_{in}$ shifts to the predicted time step) are filled with the outputs from the previous time step. Hence, input to the network is defined as

$$s(t) = (a_{\mathbf{t}_1}, \hat{a}_{\mathbf{t}_2}, u^i_{\mathbf{t}_1}, \hat{u}^i_{\mathbf{t}_2}, u^s_{\mathbf{t}_1}, \hat{u}^s_{\mathbf{t}_2}), \tag{9}$$

$$\{\mathbf{t}_1 | t - T_{in} + 1 \le \mathbf{t}_1 \le t\}, \tag{10}$$

$$\{\mathbf{t}_2 | t + 1 \le \mathbf{t}_2 \le t + (T - T_{in})\}. \tag{11}$$

As shown in Fig. 2, the prediction segment of the recurrent input is generated by shifting the corresponding previous outputs of the network to the time direction for one step.

4. Experiments on cross-modal memory retrieval and behavior recognition

4.1. *Construction of the proposed framework*

Fig. 3 depicts a schematic diagram of our proposed framework. Two independent deep neural networks are utilized for image compression and temporal sequence learning. The image compression network, shown in Fig. 3(a), inputs raw RGB color images acquired from a camera mounted on the head of the robot and outputs the corresponding feature vectors from the central hidden layer. The image features are synchronized with the joint angle vectors acquired from both arm joints and multimodal temporal segments are generated. The multimodal temporal segments are then fed into the temporal sequence learning network (i.e., Fig. 3(b)). Accordingly, multimodal features are acquired from the central hidden layer, while reconstructed multimodal temporal segments are obtained from the output layer.

The outputs from the temporal sequence learning network are used for both robot motion generation and image retrieval. The joint angle outputs from the network are rescaled and resent to the robot as joint angle commands for generating motion. The network can also reconstruct the retrieved images in the original form by decompressing the image feature outputs, because the image compression network models the identity map from the inputs to the outputs via feature vectors in the central hidden layer.

4.2. *Experimental setup*

Our proposed mechanisms are evaluated by conducting object manipulation experiments with the small humanoid robot NAO, developed by Aldebaran Robotics [32]. The multimodal data, including image frames and joint angles, are recorded synchronously at approximately 10 fps. For the image data input, the original 320 × 240 image is resized to a 20 × 15 matrix of pixels in order to meet the memory resource availability limitation of our computational environment.[1] For joint angle data input, 10 degrees of freedom of the arms (from the shoulders to the wrists) are used.

As shown in Fig. 4, six different object manipulation behaviors identified by different colorful toys are prepared for training. We record the multimodal temporal sequence data by generating different arm motions corresponding to each object manipulation by direct teaching. The resulting lengths of the motion sequences are between 100 and 200 steps, which is equivalent to between 10 and 20 s. To balance the total motion sequence lengths between different behaviors, direct teaching is repeated six to 10 times for each behavior, such that the number of repetitions becomes inversely proportional to the motion sequence length. Among all of the repetitions, one result is used as test data and the others are used as training data. For multimodal temporal sequence learning, we use a contiguous segment of 30 steps from the original time series as a single input. By sliding the time window by one step, consecutive data segments are generated.

[1] We utilized a personal computer with an Intel Core i7-3930K processor (3.2 GHz, 6 cores), 32 GB main memory, and a single Nvidia GeForce GTX 680 graphic processing unit with 4 GB on-board graphics memory. Because the size of weight matrices of a multi-layered neural network exponentially increases as the input dimension increases, we felt it sensible to keep the number of input dimensions as small as possible, as long as the dimensionality reduction did not critically degrade the quality of our experiments. As a result of preliminary experimentation, we found all of our memory retrieval experiments are feasible even with this reduced image resolution.

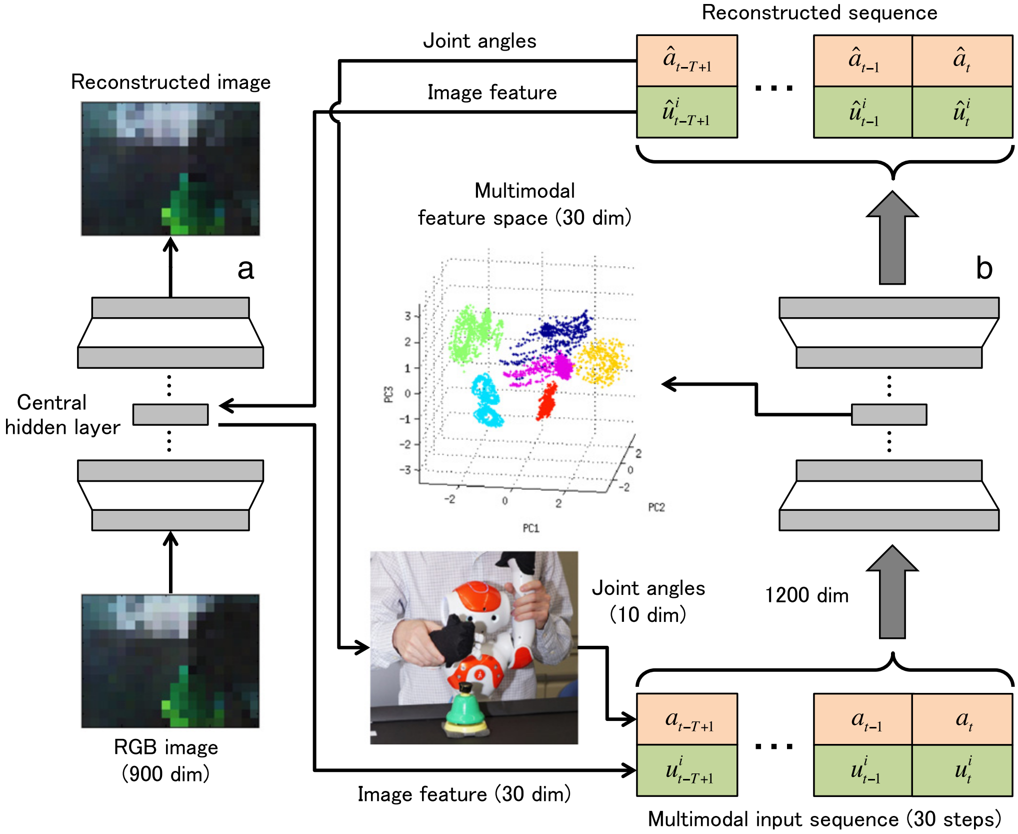
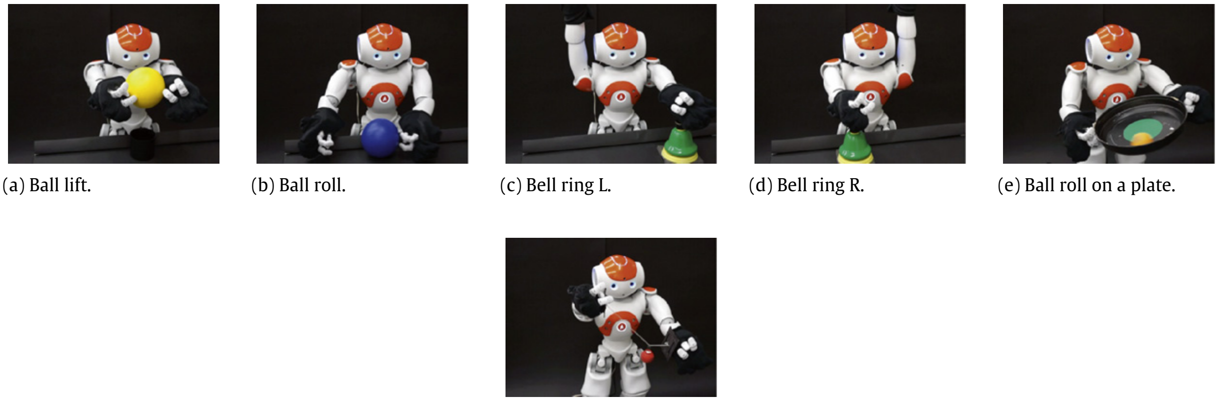


Fig. 3. Multimodal behavior learning and retrieving mechanism; two independent deep neural networks are utilized for (a) image compression and (b) temporal sequence learning.



(a) Ball lift.

(b) Ball roll.

(c) Bell ring L.

(d) Bell ring R.

(e) Ball roll on a plate.

(f) Ropeway.

Fig. 4. Object manipulation behaviors; (a) Ball lift: holding a yellow ball on the table with both hands, then raising the ball to shoulder height three times with up-and-down movements; (b) Ball roll: iteratively rolling a blue ball on top of the table to the right and left by using alternating arm movements; (c) and (d) Bell ring L/R: ringing a green bell placed on either the right or left side of the table by the corresponding arm motion; (e) Ball roll on a plate: rolling an orange ball placed in a deeply edged plate attached to both hands, and alternately swinging both arms up and down; and (f) Ropeway: swinging a red ball hanging from a string attached to both hands by alternately moving both arms up and down.

Table 1
Experimental parameters.

	TRAIN*	TEST*	I/O*	ENCODER DIMS*
IFEAT**	8444	948	900	1000-500-250-150-80-30
TSEQ**	20 548	776	1200	1000-500-250-150-80-30

* TRAIN, TEST, I/O, and ENCODER DIMS give the size of the training data, the test data, the input and output dimensions, and the encoder network architecture, respectively.
** IFEAT and TSEQ stand for image feature and temporal sequence, respectively.

Table 1 summarizes the datasets and associated experimental parameters. For both the image feature and temporal sequence learning, the same 12-layer deep neural network is used. In each case, the decoder architecture is a mirror-image of the encoder, yielding a symmetric autoencoder. The parameter settings of the network structures are empirically determined in reference to such previous studies as [27] and [33]. The input and output dimensions of the two networks are defined as follows: 900 for image feature learning, which is defined by 20 × 15 matrices of pixels for the RGB colors; and 1200 for temporal sequence learning, which is defined by the 30-step segment of the 40-dimension multimodal vector

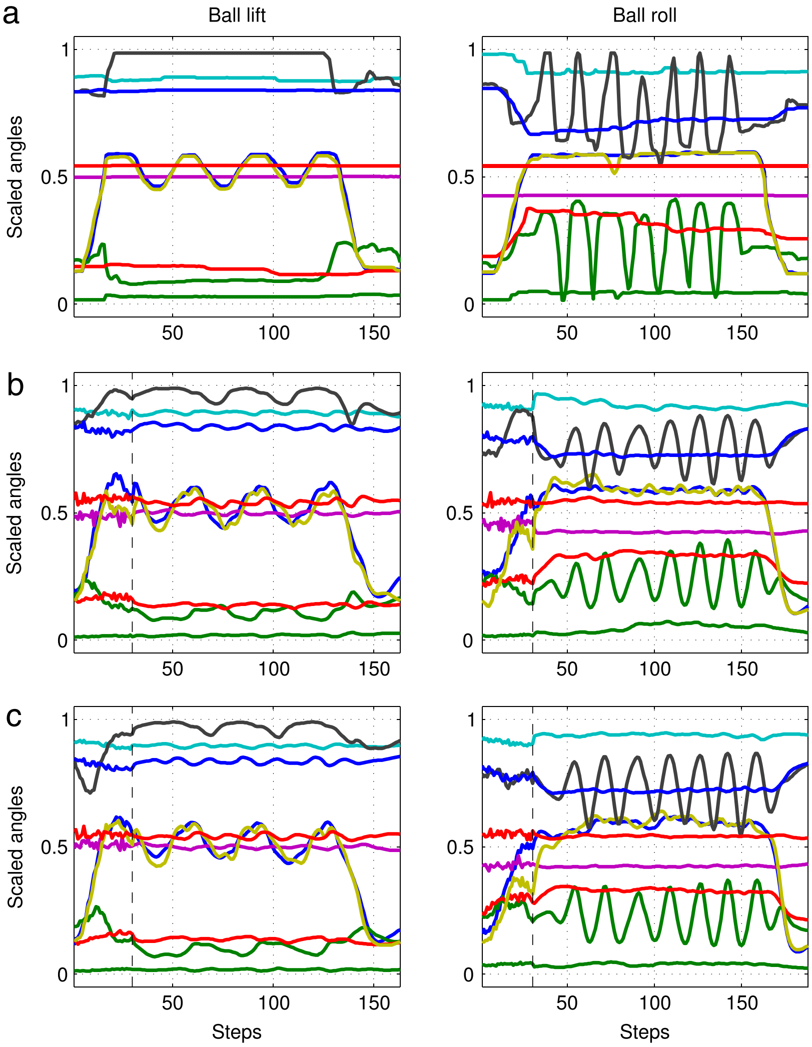


Fig. 5. Example motion reconstructions by our proposed model; graphs on the top row (a) show the original motion trajectories in the test data; graphs on the second row (b) and the bottom row (c) show the reconstructed trajectories acquired by cross-modal memory retrieval from the image sequence and temporal sequence prediction, respectively; the reconstructed trajectories correspond to the same behaviors shown on the top row.

composed of 10 joint angles and the 30-dimension image feature vector. For the activation functions, linear functions are used for both the central hidden layers and logistic functions are used for the rest of the layers in reference to [27].

The length of the time window is determined by considering the following two constraints. First, if the length of the time window increases, the network may consider longer contextual information. Second, if the length of the time window becomes too long, the dimension of the multimodal temporal vector becomes too big to be processed in an acceptable amount of time. The implicit policy is to keep the input dimensions below 3000 because of our computational limitation. As the multimodal vector dimension is 40, the temporal sequence length should be below 75. Considering the cyclic frequencies of the joint angle trajectories acquired from the six object manipulation behaviors, we determine that 30 steps are enough to characterize a phase of the behaviors.

For multimodal integration learning, we trained the temporal sequence learning network using additional examples that have only a single modality to explicitly model the correlations across the modalities [24]. In practice, we added examples that have noisy values for one of the input modalities (e.g., the image feature) and original values for the other input modality (e.g., the joint angles) but still require the network to reconstruct both modalities. Thus, one-third of the training data has only image features for input, while another one-third of the data has only joint angles and the last one-third has both image features and joint angles. For the noisy values, we superimpose Gaussian noise with a standard deviation of 0.1 on the original data.

4.3. Evaluation of cross-modal memory retrieval and temporal sequence prediction

We conducted two experiments to evaluate cross-modal memory retrieval performance. One experiment generates the joint angle sequence (motion) by providing image sequences, whereas the other generates an image sequence by providing the joint angle sequence. For these experiments, inputs to either modality of the full 30 steps are provided, and the sequence for the other modality is internally generated in a closed-loop manner (see Section 3.3). In the experiment to evaluate temporal sequence prediction, the input window length is defined as $T_{in} = 25$, and the corresponding future five steps are internally generated as predictions (see Section 3.4). For all of the experimental settings above, although the initial values for the recurrent inputs are randomly generated, the internal values eventually converge to the corresponding states in association with the input values of the other modalities by the generalization capability of the network.

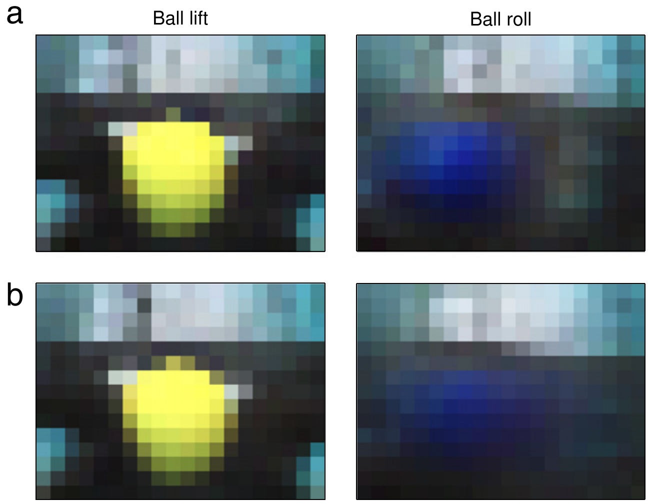


Fig. 6. Example image reconstructions by our proposed model; images on the top row (a) show the original images decompressed from the image feature vector in the test data; images on the bottom row (b) show the corresponding reconstructed images decompressed from the feature vectors acquired by cross-modal memory retrieval from the joint angle sequence.

Fig. 5 shows the example results of joint angle sequence generation from the image sequence input and temporal sequence prediction. We generated full length trajectories of the object manipulation behavior by accumulating the iteratively retrieved joint angle vectors acquired from the 30th (final) step of the temporal window. In the figure, graphs on the top row (Fig. 5(a)) are the original motion trajectories in the test data. Graphs on the second row (Fig. 5(b)) show that the appropriate trajectories are generated and the configurations of the trajectories are clearly differentiated according to the provided image sequences. Graphs on the bottom row (Fig. 5(c)) show that our proposed mechanism correctly predicted future joint angles at five steps ahead of the 25 steps of the multimodal temporal sequence. The low reconstruction qualities of the first 30 steps are attributed to the random values supplied for the recurrent inputs at the initial iteration of the generation process.

Fig. 6 shows example results of image sequence generation from the joint angle sequence input. The images shown in the figure are single frames drawn from the series of images for each behavior. Although the details of the images are slightly different, the objects showing up in the images are correctly reconstructed, and the locations of the color blobs are properly synchronized with the phases of the motion.

We conducted a quantitative evaluation of cross-modal memory retrieval by preparing 10 different initial model parameter settings for the networks and replicating the experiment of learning the same dataset composed of the six object manipulation behaviors. Table 2 summarizes these results. In the table, IMG → MTN indicates image to motion, whereas MTN → IMG indicates motion to image; further, the temporal sequence prediction (PRED) performances for the six behavior patterns are also shown. The numbers given in each entry of the table represent the root mean square (RMS) errors of the reconstructed trajectories (normalized by scaling between 0 and 1) on the test data. The RMS errors in Table 2 demonstrate that the reconstruction errors are below 10% for all of the evaluation conditions.

In detail, each of the RMS errors are calculated as

$$E_{IMG\to MTN} = \sqrt{\frac{1}{T_{seq}}\sum_{t=1}^{T_{seq}} |\tilde{a}_t - \hat{a}_t|^2}, \tag{12}$$

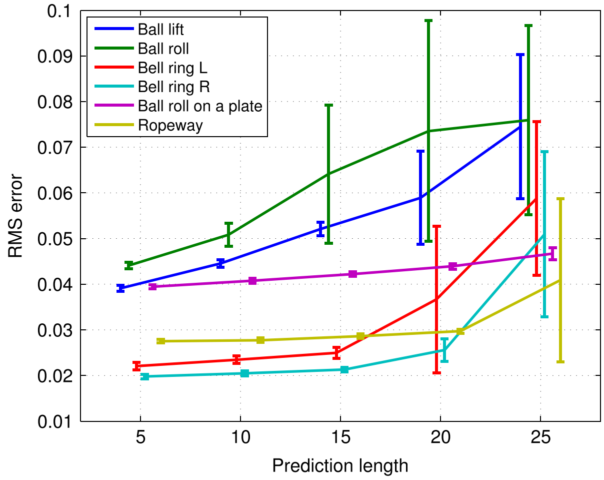


Fig. 7. Temporal sequence prediction errors of six object manipulation behaviors, depending on the prediction length; mean and standard deviation are calculated from 10 replicated learning experiments; plots are horizontally displaced from the original positions to avoid the overlap of the error bars. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$E_{MTN\to IMG} = \sqrt{\frac{1}{T_{seq}}\sum_{t=1}^{T_{seq}} |\tilde{r}^i_t - \hat{r}^i_t|^2}, \tag{13}$$

$$E_{PRED} = \sqrt{\frac{1}{T_{seq}}\sum_{t=1}^{T_{seq}} |\tilde{s}_t - \hat{s}_t|^2}, \tag{14}$$

where $E_{IMG\to MTN}$, $E_{MTN\to IMG}$, and E_{PRED} are RMS errors corresponding to the reconstruction modes identified by subscripts; $\tilde{a}_t$, $\hat{a}_t$, $\tilde{r}^i_t$, $\hat{r}^i_t$, $\tilde{s}_t$, and $\hat{s}_t$ are the truth and reconstructed vectors representing the raw image data, the joint angle, and the multimodal feature at time t, respectively; and T_{seq} is the length of the test sequence for each of the behaviors.

Finally, to analyze the temporal sequence prediction performance in more detail, we evaluated the prediction errors at the last (30th) step of the time window, depending on the prediction length, by varying the input window length T_{in} from 25 to five in decreasing steps of five. As expected, the RMS errors, as shown in Fig. 7, demonstrate that prediction error increases as prediction length increases. Nevertheless, the reconstruction errors are below 10% in all of the evaluation conditions.

4.4. Real-time adaptive behavior selection according to environmental changes

As a further experiment, we switched the robot's behavior according to changes in the objects displayed to the robot. The approach is a combination of cross-modal memory retrieval and temporal sequence prediction in the sense that the joint angles five steps ahead, considering control delay, are predicted from the previous 25 steps of the image input sequence. By iteratively sending the predicted joint angles as the target commands for each joint angle of the robot, the robot generates motion in accordance with environmental changes. For the initial trial, we tested the raw image input and confirmed that the robot can properly select behaviors according to changes of the displayed object. However, we found that the reliability of our current image feature vector is easily affected by the environmental lighting conditions.[2]

[2] We recognize that the instability of the image feature vector under real environment is due to the limitation on the variation in our image dataset utilized for training the image feature-extraction network.

Table 2
Reconstruction errors.

	IMG → MTN	MTN → IMG	PRED
LIFT*	7.11×10^{-2} (1.44×10^{-3})	1.76×10^{-2} (8.99×10^{-4})	3.91×10^{-2} (6.47×10^{-4})
ROLL*	7.05×10^{-2} (1.55×10^{-3})	4.45×10^{-2} (1.20×10^{-3})	4.41×10^{-2} (7.33×10^{-4})
RING-L*	4.95×10^{-2} (2.64×10^{-3})	1.83×10^{-2} (4.72×10^{-4})	2.21×10^{-2} (8.19×10^{-4})
RING-R*	3.64×10^{-2} (2.61×10^{-3})	1.79×10^{-2} (3.64×10^{-3})	1.98×10^{-2} (4.90×10^{-4})
PLT*	8.98×10^{-2} (1.35×10^{-3})	1.49×10^{-2} (2.96×10^{-3})	3.94×10^{-2} (4.34×10^{-4})
RWY*	5.63×10^{-2} (9.50×10^{-4})	1.89×10^{-2} (5.32×10^{-3})	2.75×10^{-2} (4.32×10^{-4})

* LIFT, ROLL, RING-L, RING-R, PLT, and RWY stand for ball lift, ball roll, bell ring L, bell ring R, ball roll on a plate, and ropeway, respectively.
** Standard deviations in parentheses.

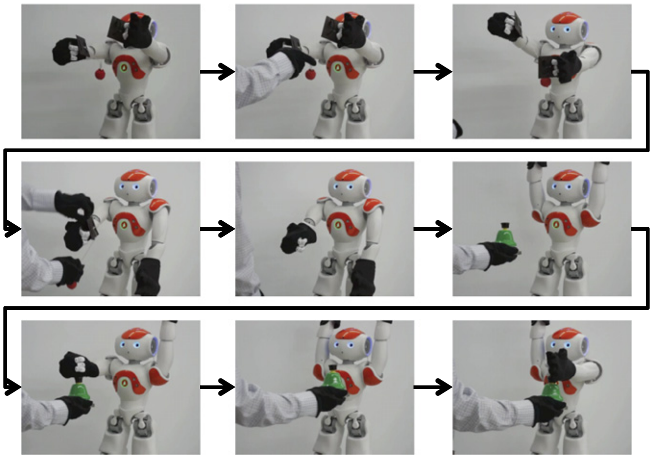


Fig. 8. Real-time transition of object manipulation behaviors according to changes in the displayed objects; behavior changes are shown in the order of Ropeway, Bell ring R, and Bell ring L.

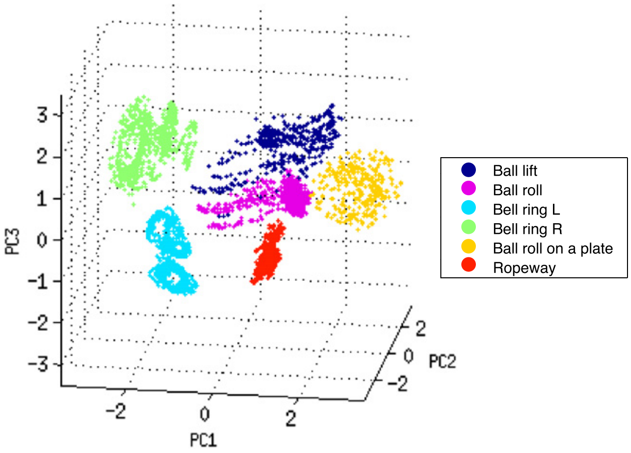


Fig. 9. Acquired multimodal feature space; PC1, PC2, and PC3 axes correspond to principal components 1, 2, and 3, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Therefore, we adopted the color region segmentation and used the coordinates of the center of gravity of the color blobs as a substitution to the image feature vector for the perception stability under various lighting conditions. As a result, we succeeded in switching multiple behaviors based on the displayed objects. Fig. 8 shows photos of the transition from one behavior to the next.

4.5. Visualization of multimodal feature space

Fig. 9 presents the scatter plot of the three-dimensional principal components of the acquired multimodal features. The multimodal feature vectors are generated by recognizing the training data from the temporal sequence learning network and recording the activations of the central hidden layer. This figure demonstrates that the feature space is segmented according to the different object manipulation behaviors and the feature vectors are self-organizing multiple clusters. The structure of the multimodal feature space suggests that a supervised discrimination learning of multiple behaviors might be possible by modeling correspondences between the acquired multimodal features and the behavior categories.

4.6. Evaluating behavior recognition performance using multimodal features

In this subsection, we examine how the acquired multimodal feature expression contributes to the robustness of a behavior recognition task. In our learning framework, raw sensory inputs are converted to sensory features, and the multiple sources of sensory features are integrated together to generate multimodal features utilizing the dimensionality compression function of an autoencoder. Making efficient use of the higher-level features, we can expect the following two effects in the behavior recognition task: (1) a discrimination model can improve its categorization performance against noisy sensory inputs by exploiting the higher generalization capabilities of the compressed representations; and (2) the integrated representation of multimodal inputs helps to inhibit the degradation of categorization performance by complementing a decrease in reliability of sensory input with information from the other modalities.

To verify our hypotheses, we evaluated the noise robustness of a behavior discrimination mechanism under different training conditions using the joint angle test sequences corresponding to the six object manipulation behaviors. More specifically, we compare the variation in behavior recognition rates depending on the differences of the standard deviation of Gaussian noise superimposed on the joint angle sequences. To investigate the effects of the higher-level features acquired from dimensionality compression and multimodal integration, we compare the performance of the classifier under the following four different training conditions:

- (1a) MTN (raw): Raw joint angles are used as inputs.
- (1b) MTN (compressed): Joint angle feature vectors are used as inputs. Feature vectors are generated by compressing the joint angle sequences utilizing an autoencoder.[3]
- (2a) MTN + IMG: Multimodal feature vectors are used as inputs. Feature vectors are generated by compressing the joint angle sequences and the corresponding image feature sequences utilizing the temporal sequence learning network. Image feature sequences are generated by compressing the clean image sequences acquired from the test data.

[3] The structure of the autoencoder used in (1b) is as same as that of the temporal sequence learning network used for the multimodal integration learning in (2a), except that the image feature inputs are excluded.

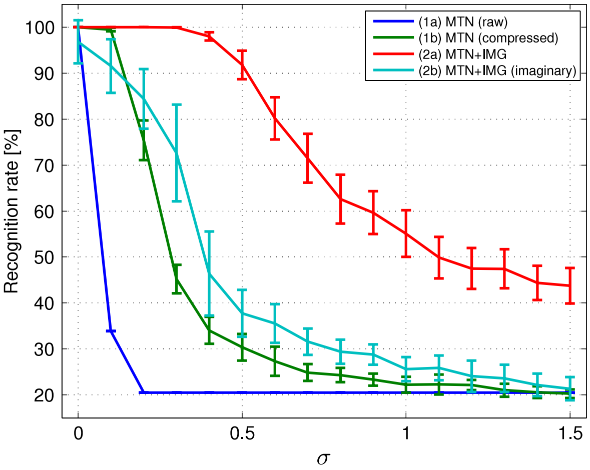


Fig. 10. Behavior recognition rates depending on the changes in standard deviation σ of the Gaussian noise superimposed on the joint angle sequences; the amplitudes of the joint angles are normalized to the range 0–1; mean and standard deviation are calculated from 10 replicated learning experiments. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

- (2b) MTN + IMG (imaginary): Multimodal feature vectors are used as inputs. In this case, the image feature sequences are self-generated inside the network instead of externally generated from the test data.

All of the training conditions, except for case (1a), are statistically evaluated on the 10 replicated learning results (see Section 4.3).

The compressed feature vector sequences are acquired by recording the activation patterns of the central middle layer of the temporal sequence network. As one of the most popular classification algorithms with an excellent generalization capability, a support vector machine (SVM) – namely, the multi-class SVM using one-against-all decomposition in the Statistical Pattern Recognition Toolbox for MATLAB [34] – is used as a classifier. An RBF kernel with default parameters (provided by the toolbox) is used to address the one-against-all multiclass non-linear separation of the acquired multimodal features; further, the Sequential Minimal Optimizer (SMO) is used as the solver for the computational efficiency.

Fig. 10 shows the variations of the behavior recognition rates depending on the changes in standard deviation of the Gaussian noise superimposed on the joint angle sequences. The results demonstrate three remarkable advantages of utilizing higher-level features for the behavior recognition task. First, comparing results of (1b) with (1a) shows the superior performance of compressed joint angle features over raw joint angles in regards to behavior recognition robustness. Second, comparing results of (2a) with (1b) shows that the multimodal features manifest higher noise robustness over single modal features by suppressing the negative effects caused by the degradation of the reliability of joint angles; this is achieved by making effective use of the complementary information from the image features. Third, comparing results of (2b) with (1b) demonstrates that even when the joint angle modality is provided as the sole input, the self-generated sequences for the image features still help to prevent degradation in behavior recognition performance. From these results, we confirmed our hypotheses that the use of higher-level features acquired by compressing raw sensory inputs and integrating multimodal sequences contribute to noise resistance of the behavior recognition tasks.

5. Experiments on intersensory causality modeling

In the previous experiments, we demonstrated that our proposed framework succeeds in cross-modal memory retrieval and stable behavior recognition utilizing the self-organized multimodal fused representations. In this section, we extend the experimental setting by incorporating sound signals as another input modality. Through our experimentation, we investigated how our proposed framework extracts the intersensory causality from the sensory-motor experience in the environment and predicts the sensory outcomes utilizing the acquired causality model.

5.1. Construction of our proposed framework

Fig. 11 shows a schematic diagram of our proposed framework. Three independent deep neural networks (i.e., autoencoders) are utilized for sound compression, image compression, and temporal sequence learning. The sound data acquired from a microphone mounted on the head of the robot is preprocessed by discrete Fourier transform (DFT). The sound compression network (Fig. 11(a)) inputs the acquired sound spectrums and outputs the corresponding feature vectors from the central hidden layer. Similarly, the image compression network (Fig. 11(b)) inputs raw RGB bitmap images acquired from a camera mounted on the head of the robot and outputs the corresponding feature vectors. The sound and image features are synchronized with the joint angle vectors, and multimodal temporal segments are generated. These multimodal temporal segments are then fed into the temporal sequence learning network (Fig. 11(c)). Accordingly, multimodal features and reconstructed multimodal segments are output from the central hidden layer and the output layer of the network, respectively.

The outputs from the temporal sequence learning network can be used for robot motion generation, sound spectrum retrieval, or image retrieval. The joint angle outputs from the network are rescaled and resent to the robot as joint angle commands for generating motion. The networks can also reconstruct the retrieved sound spectrum or images in the original form by decompressing the corresponding feature outputs because the sound compression network and the image compression network model the identity map from the inputs to the outputs via feature vectors in the central hidden layer.

5.2. Experimental setup

The cross-modal memory retrieval performance of our proposed mechanisms is evaluated by conducting bell-ringing tasks with the same robot used in our first experiment. The bell-ringing task is setup as follows: three different desktop bells, which can be identified by either the surface color or the sound pitch, are prepared for the experiment. Correspondences between the colors and the pitch notations are shown in Fig. 12(a). For each bell-ringing trial, two bells are selected and placed in front of the robot side by side. Then, either one of the two bells is rung by hitting a button on top of the bell. Due to the limited outreach of the hands, each side of the bell can be rung only with the corresponding arms. As shown in Fig. 12(b), there are six possible bell placement combinations. Note that under the task configuration, information from at least two different modalities is required to determine the right bell-ringing situation. In practice, the robot cannot (1) determine which bell is going to be rung only from the initial image, (2) determine the placement of the ringing bell only from the sound, and (3) predict what sound will come out only from the arm motion.

We record twelve different multimodal temporal sequence datasets by generating the right and left bell-striking motions under the six different bell placement configurations. Arm joint angle sequences corresponding to the bell-striking motions are generated by the angular interpolation of the initial and target postures. Pulse-code modulation (PCM) sound data is recorded with a 16 kHz sampling rate, a 16-bit depth, and a single channel

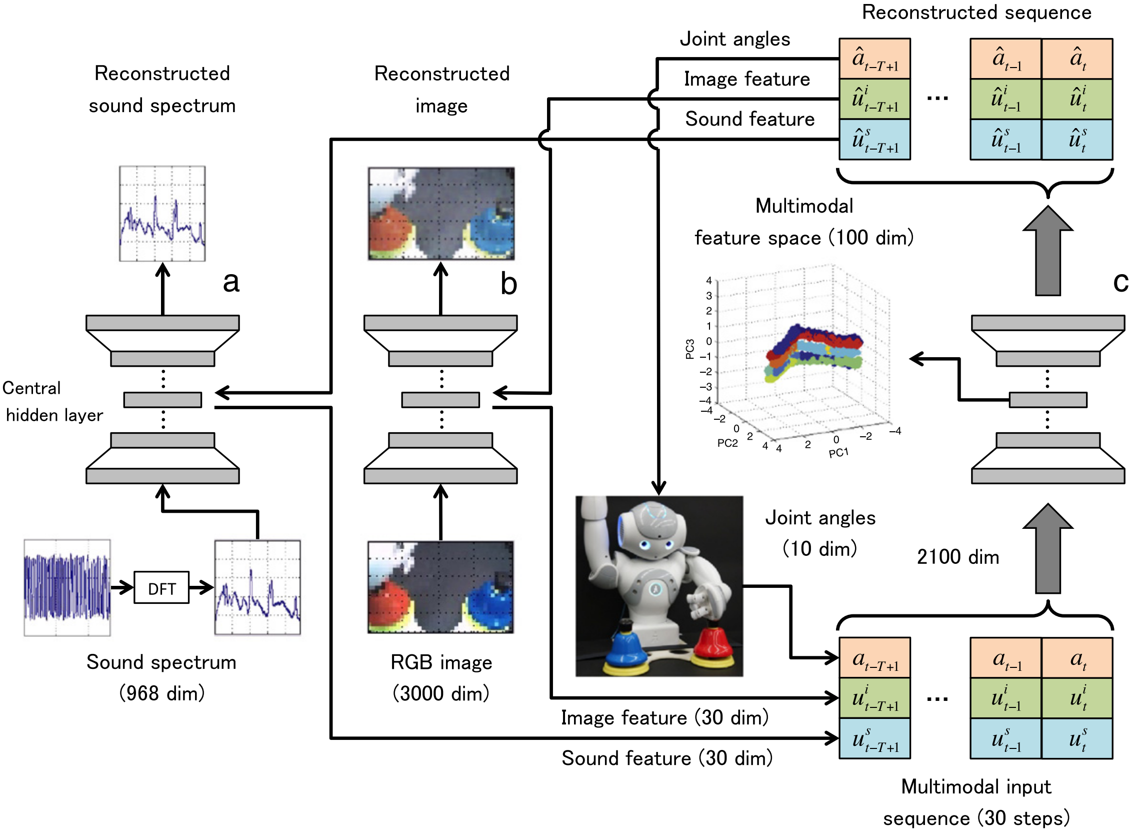
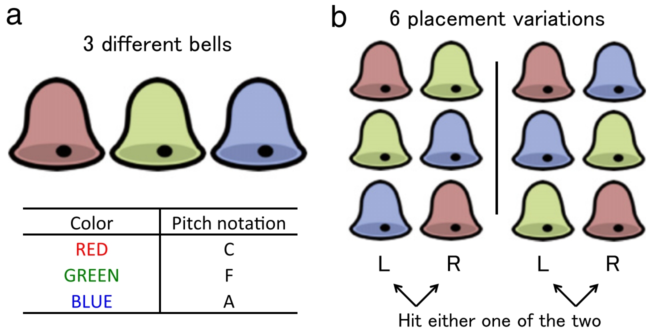

Reconstructed sequence
Joint angles
Image feature
Sound feature
$\hat{a}_{t-T+1}$ $\hat{u}^i_{t-T+1}$ $\hat{u}^s_{t-T+1}$ … $\hat{a}_{t-1}$ $\hat{a}_t$ $\hat{u}^i_{t-1}$ $\hat{u}^i_t$ $\hat{u}^s_{t-1}$ $\hat{u}^s_t$
Reconstructed sound spectrum
Reconstructed image
Multimodal feature space (100 dim)
a
b
c
Central hidden layer
PC3
PC2
PC1
Joint angles (10 dim)
2100 dim
DFT
Sound spectrum (968 dim)
RGB image (3000 dim)
Image feature (30 dim)
Sound feature (30 dim)
a_{t-T+1} u^i_{t-T+1} u^s_{t-T+1} … a_{t-1} a_t u^i_{t-1} u^i_t u^s_{t-1} u^s_t
Multimodal input sequence (30 steps)

Fig. 11. Multimodal behavior learning and retrieving mechanism; compared with the previous experimental setup shown in Fig. 3, this experimental setup incorporates another deep neural network (a) for sound feature extraction.

with a microphone mounted on the forehead of the robot.[4] The image frames and the joint angles of both arms are recorded at approximately 66 Hz, which includes replicated image frames. To synchronize the sound data with the image and joint angles data, the sound data is preprocessed by a DFT with a 242-sample hamming window and 242 samples of window shift with no overlap. A partial region of 320 × 200 pixels is cropped from the original 320 × 240 image and resized to 40 × 25 pixels to meet the memory resource availability limitation on our computational environment. For the joint angle data input, 10 degrees of freedom of the arms (from the shoulders to the wrists) are used. The resulting lengths of the motion sequence were approximately 200 steps each, which is equivalent to about 3 s each. For multimodal temporal sequence learning, we used contiguous segments of 30 steps from the original time series as a single input. By sliding the time window by one step, consecutive data segments are generated.

Table 3 summarizes the datasets and associated experimental parameters. For both the sound feature and the image feature learning, the same 12-layered deep neural networks are used. For temporal sequence learning, a 10-layer network is used. In each case, the decoder architecture is a mirror-image of the encoder, yielding a symmetric autoencoder. The parameters for the network structures are empirically determined in reference to such previous studies as [27] and [33]. The input and output dimensions of the two networks are defined as follows: 968 for sound feature learning, which is defined by binding consecutive four-step sequences of the sound spectrums (i.e., 242 dimension) into a single vector, 3000 for the image feature learning, which is defined by 40 × 25 matrices of pixels for RGB colors, and 2100 for temporal sequence learning, which is defined by the 30-step segment of the 70-dimensional multimodal vector composed of a 30-dimensional sound feature vector, a 30-dimensional image feature vector, and 10 joint angles. Especially for the central hidden layer of the temporal sequence learning network, we compared several numbers of nodes, i.e., 30, 50, 70, and 100. By evaluating the performance of image retrieval from the sound and joint angle inputs, we concluded that 100 nodes are needed to achieve the desired memory reconstruction precision. For the activation functions, linear functions are used for all of the central hidden layers, and logistic functions are used for the rest of the layers in reference to [27].



Color	Pitch notation
RED	C
GREEN	F
BLUE	A

Fig. 12. Bell placement configurations of the bell-ringing task. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

[4] Because of the physical structure of the robot, the microphone is located close to both of the arms, which are utilized to hit the bells. Therefore, the actuation sounds of the geared reducers equipped to the arm joints are inevitably recorded in addition to the bell sounds. To avoid the degradation of memory retrieval performance arising from the actuation sounds, we introduced a brief pause to the bell hitting motion when the hand contacted the button on top of the bell.

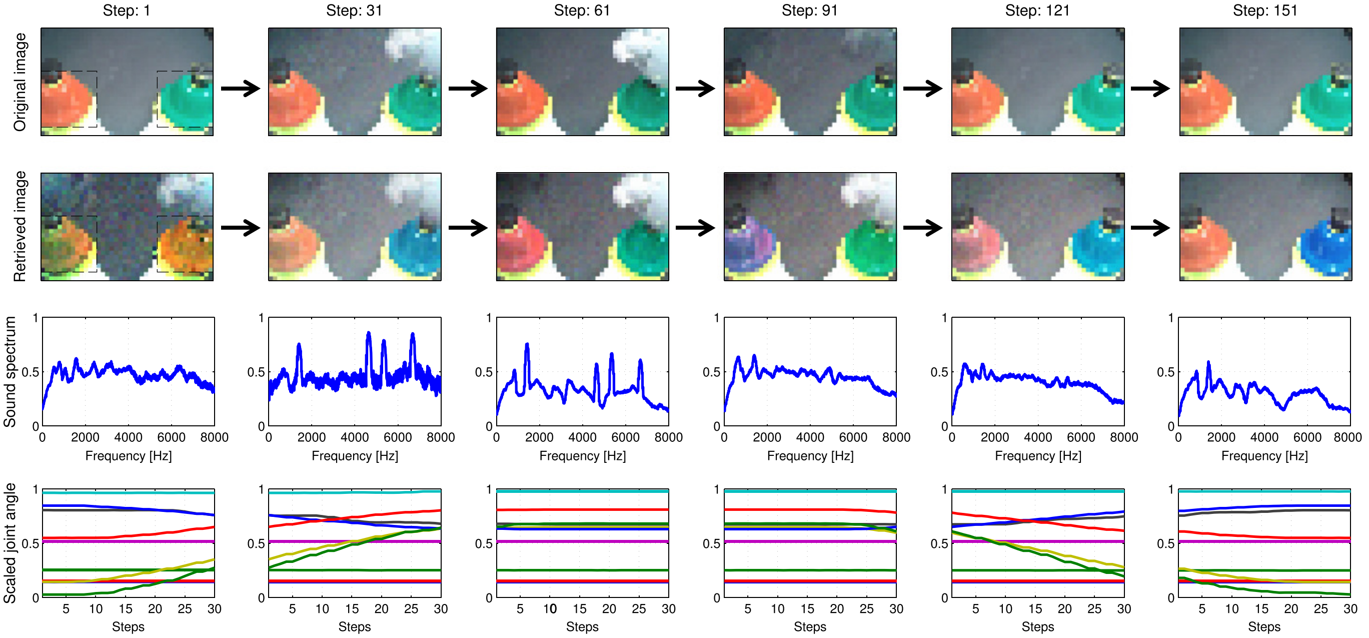


Fig. 13. Example image retrieval results from the sound and joint angles inputs; the top row and the second row show the original and retrieved images, respectively; the third and the bottom row show the sound spectrums and the 30 previous steps of joint angles sequences used as inputs to the temporal sequence learning network to retrieve the corresponding images; black dashed squares in the images at step 1 indicate the bell image regions used for image retrieval performance evaluation. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Experimental parameters.

	TRAIN*	I/O*	ENCODER DIMS*
SFEAT**	5352	968	1000-500-250-150-80-30
IFEAT**	2688	3000	1000-500-250-150-80-30
TSEQ**	8736	2100	1000-500-250-150-100

* TRAIN, I/O, and ENCODER DIMS give the size of the training data, the input and output dimensions, and the encoder network architecture, respectively.
** SFEAT, IFEAT, and TSEQ stand for sound feature, image feature, and temporal sequence, respectively.

5.3. Image sequence retrieval from sound and motion sequences

We conducted an evaluation experiment of the cross-modal memory retrieval performance by generating image sequences from the sound and joint angle input sequences. Note that in the following results, the number of sequence steps indicates the generation step rather than the recorded data step. More specifically, data from 29 steps before the beginning of the generation step are used for acquiring the initial step of the generated sequence.

Fig. 13 shows an example of image generation results from the sound and joint angle inputs. At step 1, the bells in the retrieved image are arbitrarily colored, because the color of the placed bell is not derivable before acquiring any sound input. By contrast, the image of the robot's right hand is already included in the retrieved image, because the joint angles input data indicate that the right arm is going to be used for striking the bell. At steps 31 and 61, the bell is rung, and the corresponding sound spectrum is acquired. Then, the task configuration becomes evident, and the information that the rung bell on the right side has the pitch 'F' is correlated with the color green. Thus, the color of the right bell in the retrieved image changes from the randomly initialized one to green by associating the sound and joint angles information. Conversely, the color of the left bell in the retrieved image is not stable during the run because no information is acquired from the sound input for identifying which bell is placed on the left side. Nevertheless, the retrieved image shows that when the color of the rung bell (i.e., green) is identified, the color of the other bell is selected from the remaining two colors (i.e., red or blue). This result reflects the current task design in which the color of the two bells is always different.

From step 91 (or so), the sound of the bell starts to decay, and the actuation noise of the manipulator by the posture initialization becomes dominant. Thus, the colors of the bells again become arbitrary.

5.4. Quantitative evaluation of image retrieval performance

We conducted an evaluation experiment to quantitatively examine whether our proposed model has succeeded in modeling the causality between the image, sound, and motion modalities. We prepared 10 different initial model parameter settings for the networks and replicated the experiment of learning the same dataset composed of the 12 combinations of the bell placements and bell-striking motion patterns. As a result of cross-modal image retrieval for the 10 learning results, 120 patterns of the image sequences were acquired. Image retrieval performance is quantified by the root mean square (RMS) errors of the manually selected left and right bell regions in the retrieved image (which are 13 × 13 pixels each, as indicated in Fig. 13) against the corresponding regions of the original image.

Fig. 14 shows the time variation of the image retrieval error displayed in association with the maximum value of the sound power spectrum and the joint angles sequence. The evaluation results demonstrate that the image retrieval error of the left bell becomes smaller than that of the right bell when the left bell is rung, and vice versa. The time variation of the error trajectory shows that the retrieval error decreases after the sound of the bell is acquired.

The shape of the error trajectory is not symmetric between the two graphs when the left bell or the right bell is rung. When the left bell is rung, the image retrieval error for the left bell maintains its value even after arm posture initialization. Conversely, when the right bell is rung, the image retrieval error for the right bell increases after arm posture initialization. These differ primarily because of the asymmetry of the arm actuator noise. Due to the

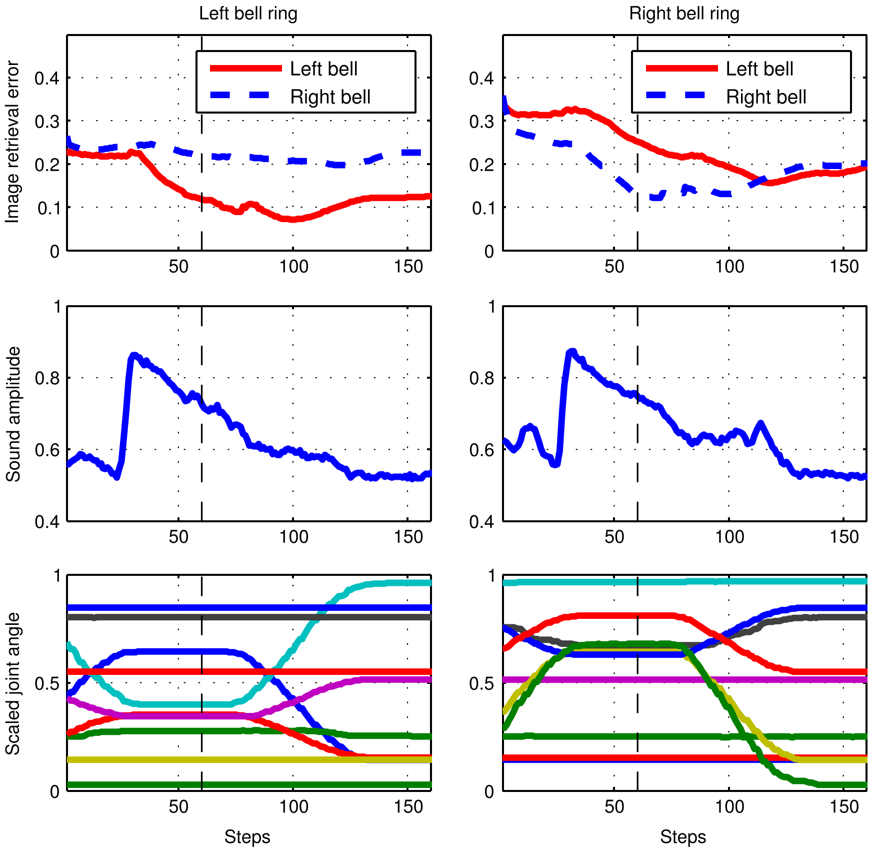


Fig. 14. Bell image retrieval errors; the graphs on the top row show the mean of the image retrieval errors from the replicated learning results (each line is acquired from 30 result sequences); the graphs on the second row show the mean of maximum sound power spectrums; and the graphs on the bottom row show the joint angles command sequences used for generating the bell-striking motion; black dashed lines indicate the time step used for evaluating the significance of the image retrieval error difference.

difference in the mechanics of the left and right actuators, which is beyond our control, the right arm produces more sound than the left arm. Hence, when the right arm posture is initialized after striking the bell, the accompanying actuator noise disturbs the internal state of the network (i.e., the data buffered in the recurrent loop), and the retrieved image is altered.

5.5. The correlation between generated motion and retrieved bell images

To evaluate the significance of the difference between image retrieval performance of the left and right regions in the same image, we conduct a t-test for the image retrieval errors at step 60 of the sequences. At that time step, the arm is brought down and the hand stably contacts the button on top of the bell. Therefore, there is no influence of actuation noise on image retrieval. As shown in Fig. 15, evaluation results show that the differences of the image retrieval errors between the two regions are statistically significant in both the right and left bell-ringing cases. Results further show that the spatial correlation between the bell region in the image and the physical motion is correctly modeled, as are the associations between the colors and sounds of the bells. Thus, the acquired causality model between the image, sound, and motion modalities is utilized for image retrieval.

5.6. Visualization of multimodal feature space

Finally, we conducted an analysis of the multimodal feature space acquired by the temporal sequence learning network. Among the 10 replicated learning results, we took a single result and recorded the activation patterns of the central hidden layer of the network when the 12 patterns of bell-ringing sequences are the input. We applied principal component analysis (PCA) to project the resulting 100-dimension feature vector sequences to a three-dimensional space defined by the acquired principal components. Fig. 16(a) demonstrates that the robot's motion pattern is represented in the two-dimensional space composed of the first and second principal components. In addition, Fig. 16(b) shows that the bell placement configurations are structured along the coordinate defined by the third principal component. Results of this analysis demonstrate that the causal dependencies between the multiple modalities are self-organized in the temporal sequence learning network.

6. Discussion

6.1. Information complementation capability as a means for cross-modal memory retrieval

In this study, we demonstrated the significant scalability of the deep learning algorithm applied to the time-delayed autoencoder on sensory-motor coordination learning problems. We presented experimental results on cross-modal memory retrieval and the subsequent adaptive behavior generation of a humanoid robot in the real environment. For example, in the image sequence retrieval experiment of the object manipulation behaviors learning task in 4.3, 900 dimensions of the image feature vector sequence were recalled only from the 300 dimensions of joint angle sequence input. This result shows that three times as much information was recalled by the generalization capability of the autoencoder.

This powerful information complementation capability is one of the advantages of our proposed time-delay autoencoder. Utilizing

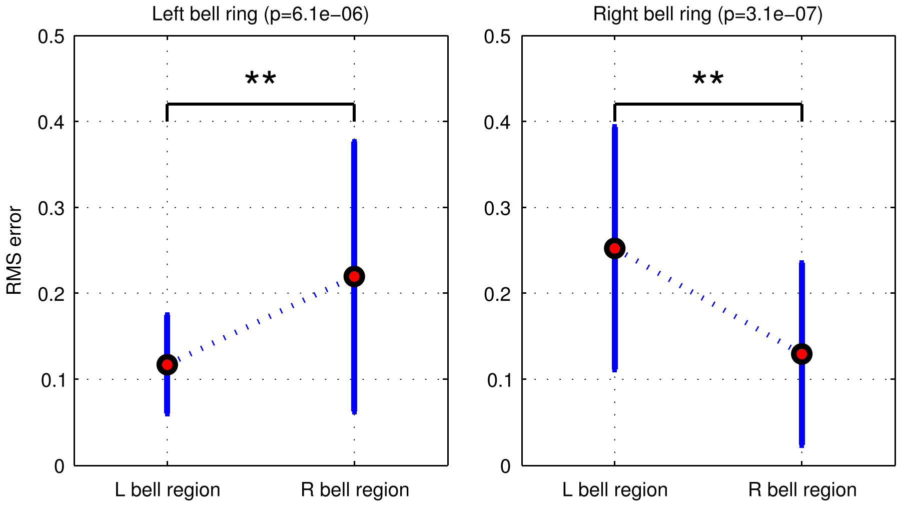


Fig. 15. Bell image retrieval errors at step 60; red circles and blue bars denote the mean and standard deviation of the errors from 10 replicated learning experiments, respectively; a p value less than 0.01 is considered statistically significant (**: $p < 0.01$).

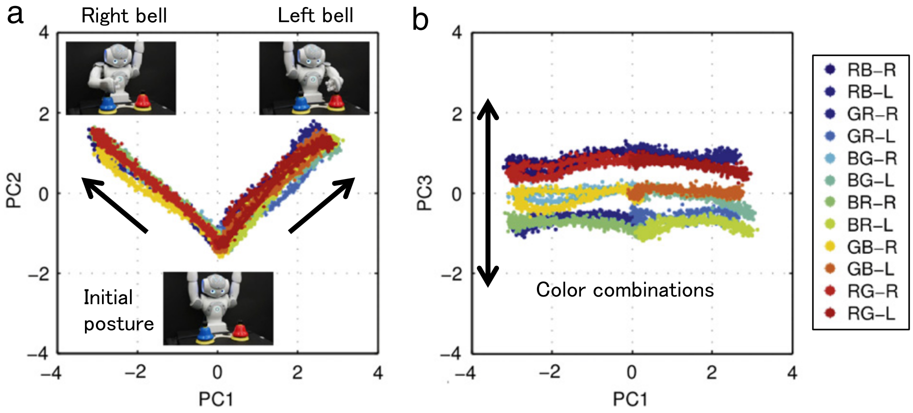


Fig. 16. Multimodal feature space and the correspondence between the coordinates and modal-dependent characteristics; the graph on the left side (a) shows the abstracted feature spaces defined by the first and second principal components, whereas the graph on the right side (b) shows the abstracted feature spaces defined by the first and third principal components; the abbreviations in the legend box indicate the color combinations of the placed bells, followed by the position (R or L) of the rung bell.

the generalization capability of the preceding half layers of the autoencoder, higher-level features that represent specific object manipulation behavior can be generated even from partial modal inputs. Further, as the autoencoder can reconstruct the original inputs from the feature vector, the predicted outputs can be recursively fed back to the input nodes, and the inputs can be used as a substitution for any lacking modality information. This recursive information loop in our proposed framework enabled a high level of stability in cross-modal memory retrieval performance.

The number of layers and the number of nodes are important factors to explain the memory capacity and generalization capability of a deep neural network; however, in general, a clear explanation has not been made for the correlation between the network structure and its learning capability. Thus, the design principle on the structure of neural networks has little theoretical foundation at the moment. This might be an important research topic for future consideration.

6.2. *Three factors that contribute to robustness in behavior recognition tasks*

Our experimental results regarding behavior recognition evaluation demonstrated that the compressed temporal features enable robust recognition performance. By comparing the recognition rates from the different evaluation conditions, we have shown that the following three factors contribute to noise robustness in behavior recognition tasks: (1) utilization of higher-level features; (2) utilization of multimodal information; and (3) utilization of self-generated sequences in multimodal behavior recognition. Below, we present our views regarding the functions of the three factors in relation to the internal mechanisms of our proposed framework.

6.2.1. *Utilization of higher-level features*

In previous work, Le et al. showed that it is possible to train neurons to be selective for high-level concepts using entirely unlabeled data [19]. As a practical result, they succeeded in acquiring class-specific neurons such as cat and human body detector neurons by training deep neural networks with unlabeled YouTube datasets. This result – i.e., that meaningful features can be self-organized even with unlabeled data – demonstrates the advantage of utilizing an autoencoder as a feature-extraction mechanism. Comparable results are presented in related works involving image classification tasks [21] and speech recognition tasks [20]. Considering all of these previous studies, our behavior recognition results seem to coincide with the view that deep neural networks produce higher-level features that have a prominent generalization capability by accumulating many layers of nonlinear feature detectors to progressively represent complex statistical structure in the data [35].

6.2.2. Utilization of multimodal information

From the viewpoint of the amount of information acquired from the multimodal sequence, the multimodal temporal sequence learning network has a clear advantage in generating a more accurate internal representation than that of a unimodal temporal sequence learning network. This fact is presented in the behavior recognition results with noisy joint angle inputs and clear image inputs from the training dataset (i.e., (2a) MTN+IMG) in 4.6. These results demonstrate that even after the joint angle information becomes uninformative, the degradation of the recognition rate converges to a level that surpasses other results. In this case, the clear image feature inputs served as a source of information for the higher-level features to correctly represent the behavior category against the uninformative joint angle inputs. Current results on the effects of multimodal learning toward robustness in recognition tasks can be regarded in the same light as, for example, improvements in multimodal speech recognition tasks utilizing the combination of sound and image inputs [24].

6.2.3. Utilization of self-generated sequences in multimodal behavior recognition

Among our behavior recognition evaluation results presented in 4.6, the most notable outcome is that a higher recognition performance is realized by the multimodal memory even with the single modal input for the joint angles (i.e., (2b) MTN+IMG (imaginary)). This result could be explained as follows. Utilizing a multimodal memory, a multimodal internal representation is generated even from noisy joint angle inputs, and successively accompanying image features are retrieved from the output nodes. As the image feature vector is recalled from the internal representation, the information becomes even more independent of the disturbance superimposed over the joint angle observations. By feeding back the retrieved image feature to the input nodes, this procedure leads to clarifying the internal representation that is equivalent to the multimodal recognition process by explicitly providing the image feature sequence to the network in parallel with the noisy joint angle sequence. In recent neuropsychological studies, the positive effects of self-referential strategies in improving memory in memory-impaired populations have been reported [36,37]. In future work, it would be interesting to further investigate how our current self-generating imaginary sequence mechanism corresponds to such psychological phenomena in the human cognitive process.

6.3. Cross-modal causality modeling as a means for sensory-motor prediction

In this study, we presented our proposed framework as able to extract implicit synchronicity among multiple modalities by integrating multimodal information. Further, the retrieved images in the bell-ringing task demonstrated that our proposed framework not only deterministically retrieved a bell image reflecting the acquired causality, but also somehow generated alternate information by selecting a candidate among multiple possibilities even if the specific situation is not identifiable for the other bell. Thus, we believe that our proposed mechanism can be utilized as a prediction mechanism for robots to infer the successive consequences of sensory-motor situations. In cognitive science studies, a sense of agency is known to be a product of the general determination of causality between action and effect, and experimental results suggest that the sense of agency arises when there is temporal contiguity and content consistency between signals related to action and those related to the putative effect [38,39,1]. Further, a recent study has reported the importance of action–effect grouping on the production of a sense of agency [4]. In all of these studies, the evaluation of spatiotemporal congruity between predicted and actual sensory feedback is considered to play an important role in the sense of agency. From our current results, we consider that our cross-modal causality modeling and subsequent memory retrieval capabilities can be utilized as a practical computational framework for sensory feedback prediction. Hence, we believe that our presented framework can be utilized in future work to promote a deeper understanding of the sense of agency.

6.4. Functional characteristics of temporal sequence learning by a time-delay autoencoder

6.4.1. Difference between our proposed time-delay autoencoder and the original time-delay neural network

The temporal sequence learning mechanism proposed in our work inputs a fixed length of time series acquired by cropping a segment of a temporal sequence within a time window. This approach inherits the idea from the original work of time-delay neural networks by Lang et al. [29]. The difference here is that the vectors identical to the inputs define the target outputs of our proposed model, whereas the symbol labels define the outputs of the original model. Consequently, one of the characteristics of our proposed model is that the compressed representation of temporal sequences is self-organized by the autoencoder, and the network can self-generate temporal sequences by recursively feeding back outputs to input nodes. The advantages of the internal sequence generation were shown by the adaptive behavior selection capability utilizing cross-modal memory retrieval and the robust behavior recognition capability with unreliable joint angle observations.

6.4.2. Characteristics of the internal representation of the temporal sequence learning network

The temporal sequence learning network is virtually modeling the dynamics of long temporal sequences by accumulatively memorizing multiple phase-wise temporal segments. Thus, a feature vector generated from a one-shot input represents a temporal phase of a sequence. This phenomenon can be confirmed from plots of the feature vectors of the bell-ringing task by observing where they formed closed loop shapes in Fig. 9. The same phenomenon can be confirmed from the second task in that the reciprocal transition of the feature vector plots on the two distinct lines corresponds to each of the right and left arm motion patterns shown in Fig. 16.

6.4.3. Length of contextual information that a time-delay autoencoder handles

The length of the input temporal segment defines the length of the contextual information handled by the temporal sequence learning network. Hence, in principle, context information longer than the temporal segment is not considered. In comparison with the other temporal sequence learning mechanisms, such as recurrent neural networks [40], this is a fundamental difference. Our proposed framework worked successfully in our experiments despite this limitation of contextual representation because the execution of robot behaviors in our task settings did not require comprehending long contextual situations. For example, for the object manipulation and bell-ringing behaviors, most of the contextual information is embedded in the environment (e.g., the robot's arm posture, position of the balls, etc.). Thus, an internal neuronal representation of the context was not required for achieving the tasks.

7. Conclusion

In this study, we proposed a deep neural network framework that enables multimodal integration learning of temporal sequences, including visual, auditory, and motion. The performance of our proposed framework was evaluated by two tasks utilizing a humanoid robot in a real-world environment. The tasks consisted of object manipulation and bell-ringing tasks. Our results demonstrated the scalability of our proposed framework in handling large amounts of training data with significant dimensionality. We presented three applications of the acquired sensory-motor integration model. First, cross-modal memory retrieval was realized. Utilizing the generalization capability of the deep autoencoder, our proposed framework succeeded in retrieving temporal sequences bidirectionally between image and motion. Second, robust behavior recognition was realized by utilizing the acquired multimodal features as inputs to supervised behavior classification learning. Third, multimodal causality modeling was realized. Our experimental results demonstrated that our proposed framework could model synchronicity between the color, pitch, and position of the bell and the corresponding bell-ringing motion from the robot's bell-ringing behaviors and memorize their correlations.

Results from the real-time transition of object manipulation behaviors in a real-world environment also revealed that our current approach for utilizing raw image data is still not stable enough for handling drastic changes in lighting conditions. Future work includes improving the robustness of the image recognition capabilities by drawing out the potential of the generalization capabilities of deep networks via the introduction of convolution networks trained with more diverse datasets. As for the current bell-ringing task, we evaluated the image retrieval performance from sound and motion with only two bell positions. In future work, it might be interesting to model a generalized representation of bell positions by training our system with bell-ringing behaviors using more variations of bell positions. Another important challenge is dynamically combining multiple sensory modalities by taking into account the relative reliability of different sensory sources. If reliability-dependent integration is attained in our framework, higher-level features might be acquired by intentionally suppressing the effects degraded modalities have on the internal representation; this might result in more robust behavior recognition performance.

Acknowledgments

This work has been supported by JST PRESTO "Information Environment and Humans" and MEXT Grant-in-Aid for Scientific Research on Innovative Areas "Constructive Developmental Science" (24119003).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.robot.2014.03.003.

References

[1] M.O. Ernst, H.H. Bülthoff, Merging the senses into a robust percept, Trends Cogn. Sci. 8 (4) (2004) 162–169.

[2] B.E. Stein, M.A. Meredith, The Merging of the Senses, The MIT Press, 1993.

[3] I. Gallagher, Philosophical conceptions of the self: implications for cognitive science, Trends Cogn. Sci. 4 (1) (2000) 14–21.

[4] T. Kawabe, W. Roseboom, S. Nishida, The sense of agency is action-effect causality perception based on cross-modal grouping, Proc. R. Soc. B: Biological Sciences 280 (1763) (2013) 20130991.

[5] R.A. Brooks, C.B. (Ferrell), R. Irie, C.C. Kemp, M. Marjanovic, B. Scassellati, M.M. Williamson, Alternative essences of intelligence, in: Proceedings of the 15th National Conference on Artificial Intelligence, Madison, Wisconsin, USA, 1998, pp. 961–968.

[6] M. Coen, Multimodal integration-a biological view, in: Proceedings of the 17th International Joint Conference on Artificial Intelligence, vol. 2, Seattle, Washington, USA, 2001, pp. 1417–1424.

[7] A. Pitti, A. Blanchard, M. Cardinaux, P. Gaussier, Distinct mechanisms for multimodal integration and unimodal representation in spatial development, in: Proceedings of the IEEE International Conference on Development and Learning and Epigenetic Robotics, San Diego, California, USA, 2012, pp. 1–6.

[8] A. Jauffret, N. Cuperlier, P. Gaussier, P. Tarroux, Multimodal integration of visual place cells and grid cells for navigation tasks of a real robot, in: Proceedings of the 12th International Conference on Simulation of Adaptive Behavior, vol. 7426, Odense, Denmark, 2012, pp. 136–145.

[9] E. Sauser, A. Billard, Biologically inspired multimodal integration: interferences in a human–robot interaction game, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Beijing, China, 2006, pp. 5619–5624.

[10] R.R. Murphy, Introduction to AI Robotics, The MIT Press, 2000.

[11] M. Ogino, H. Toichi, Y. Yoshikawa, M. Asada, Interaction rule learning with a human partner based on an imitation faculty with a simple visuo-motor mapping, Robot. Auton. Syst. 54 (5) (2006) 414–418.

[12] T. Kuriyama, T. Shibuya, T. Harada, Y. Kuniyoshi, Learning interaction rules through compression of sensori-motor causality space, in: Proceedings of the 10th International Conference on Epigenetic Robotics, Örenäs Slott, Sweden, 2010, pp. 57–64.

[13] Y. Sakagami, R. Watanabe, C. Aoyama, S. Matsunaga, N. Higaki, K. Fujimura, The intelligent ASIMO: system overview and integration, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and System, vol. 3, Lausanne, Switzerland, 2002, pp. 2478–2483.

[14] K. Kaneko, F. Kanehiro, S. Kajita, H. Hirukawa, T. Kawasaki, M. Hirata, K. Akachi, T. Isozumi, Humanoid robot HRP-2, in: Proceedings of the IEEE International Conference on Robotics and Automation, vol. 2, Barcelona, Spain, 2004, pp. 1083–1090.

[15] Willow Grarage, Personal Robot 2 (PR2), http://www.willowgarage.com/pages/pr2/overview.

[16] R. Bekkerman, M. Bilenko, J. Langford (Eds.), Scaling up Machine Learning: Parallel and Distributed Approaches, Cambridge University Press, 2011.

[17] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (1) (2009) 1–127.

[18] I. Sutskever, J. Martens, G. Hinton, Generating text with recurrent neural networks, in: Proceedings of the 28th International Conference on Machine Learning, Bellevue, Washington, USA, 2011, pp. 1017–1024.

[19] Q.V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G.S. Corrado, J. Dean, A.Y. Ng, Building high-level features using large scale unsupervised learning, in: Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, 2012, pp. 81–88.

[20] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition, IEEE Signal Process. Mag. 29 (2012) 82–97.

[21] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems 25, Lake Tahoe, Nevada, USA, 2012, pp. 1106–1114.

[22] Rosenberg Chuck, Improving photo search: a step across the semantic gap, http://googleresearch.blogspot.jp/2013/06/improving-photo-search-step-across.html (Jun. 2013).

[23] Hof Robert, Meet the guy who helped google beat Apple's Siri, http://www.forbes.com/sites/roberthof/2013/05/01/meet-the-guy-who-helped-google-beat-apples-siri/ (May 2013).

[24] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: Proceedings of the 28th International Conference on Machine Learning, Bellevue, Washington, USA, 2011, pp. 689–696.

[25] N. Srivastava, R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, in: Proceedings of the Advances in Neural Information Processing Systems 25, Lake Tahoe, Nevada, USA, 2012, pp. 2231–2239.

[26] J. Dewey, The reflex arc concept in psychology, Psychol. Rev. 3 (1896) 357–370.

[27] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[28] J. Martens, Deep learning via Hessian-free optimization, in: Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 2010, pp. 735–742.

[29] K. Lang, A. Waibel, G. Hinton, A time-delay neural network architecture for isolated word recognition, Neural Netw. 3 (1990) 23–43.

[30] B. Pearlmutter, Fast exact multiplication by the Hessian, Neural Comput. 6 (1) (1994) 147–160.

[31] N.N. Schraudolph, Fast curvature matrix–vector products for second-order gradient descent, Neural Comput. 14 (7) (2002) 1723–1738.

[32] Aldebaran Robotics, NAO Humanoid, http://www.aldebaran-robotics.com/Downloads/Download-document/192-Datasheet-NAO-Humanoid.html (Nov. 2012).

[33] A. Krizhevsky, G.E. Hinton, Using very deep autoencoders for content-based image retrieval, in: Proceedings of the 19th European Symposium on Artificial Neural Networks, Bruges, Belgium, 2011.

[34] V. Franc, V. Hlavac, Statistical pattern recognition toolbox for matlab, http://cmp.felk.cvut.cz/cmp/software/stprtool/ (Aug. 2008).

[35] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

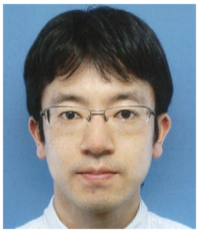
[36] M.D. Grilli, E.L. Glisky, The self-imagination effect: benefits of a self-referential encoding strategy on cued recall in memory-impaired individuals with neurological damage, JINS 17 (5) (2011) 929–933.

[37] M.D. Grilli, E.L. Glisky, Self-imagining enhances recognition memory in memory-impaired individuals with neurological damage, Neuropsychology 24 (6) (2010) 698–710.

[38] A. Pouget, S. Deneve, J. Duhamel, A computational perspective on the neural basis of multisensory spatial representations, Nature Rev. Neurosci. 3 (2002) 741–747.

[39] S. Deneve, A. Pouget, Bayesian multisensory integration and cross-modal spatial links, J. Physiol. Paris 98 (1–3) (2004) 249–258.

[40] J. Martens, I. Sutskever, Learning recurrent neural networks with Hessian-free optimization, in: Proceedings of the 28th International Conference on Machine Learning, Bellevue, Washington, USA, 2011, pp. 1033–1040.



Kuniaki Noda received the B.S. and M.S. in Mechanical Engineering in 2000 and 2002, respectively, from Waseda University, Japan. From 2002 to 2012, he worked for Sony Corporation. From 2009 to 2010, he was a visiting researcher at EPFL, Switzerland. Currently, he is a Ph.D. candidate at Waseda University. His research interests include autonomous robot, multimodal integration, deep learning, and high performance computing on GPU. He received various awards including the Hatakeyama Award from the Japan Society of Mechanical Engineers in 1999, the Best Paper Award of ICDL-EPIROB 2011, and the Best Paper Award of RSJ in 2012.

Hiroaki Arie received the B.E., M.E. and Ph.D. degrees in Engineering from Waseda University, Japan, in 2003, 2005 and 2011 respectively. From 2008 to 2011, he was a Research Associate in the Laboratory for Behavior and Dynamic Cognition, Brain Science Institute, RIKEN. From 2011 to 2012, he was a Research Scientist at the Brain Science Institute, RIKEN. He is currently an Assistant Professor at the Department of Intermedia Art and Science, Waseda University. His interests include neuroscience, complex adaptive system and developmental robotics.

Yuki Suga received the Dr. of Engineering degree in 2008, in School of Mechanical Engineering, Waseda University. From 2009 to 2011, he worked for Revast Co., Ltd. Since 2012, he is an owner engineer of the Sugar Sweet Robotics Co., Ltd. Since 2009, he has been a visiting research associate of Humanoid Robot Institute, Waseda University. Since 2012, he has been a visiting research associate of Shibaura Institute of Technology. Since 2012, he has been a project researcher of Graduate School of Information Science and Technology, the University of Tokyo. He is a member of Robotics Society of Japan (RSJ). His research interests are human–robot interaction and the robotics software platforms.

Tetsuya Ogata received the B.S., M.S. and D.E. degrees in Mechanical Engineering, in 1993, 1995 and 2000, respectively, from Waseda University. He was a Research Fellow of JSPS, a Research Associate of Waseda University, a Research Scientist of RIKEN Brain Science Institute, and an Associate Professor of Kyoto University. He is currently a Professor of Faculty of Science and Engineering, Waseda University. Since 2009, he has been a JST PRESTO researcher. His research interests include human–robot interaction, dynamics of human–robot mutual adaptation and inter-sensory translation in robot systems. He is a member of IEEE, RSJ, JSAI, IPSJ, JSME, SICE, etc.