Pattern Recognition Letters 31 (2010) 1560–1569

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Inter-modality mapping in robot with recurrent neural network

Tetsuya Ogata [a,*], Shun Nishide [a], Hideki Kozima [b], Kazunori Komatani [a], Hiroshi G. Okuno [a]

[a] Graduate School of Informatics, Kyoto University, Yoshida-honmachi Sakyo-ku, 606-8501 Kyoto, Japan
[b] School of Project Design, Miyagi University, 1 Gakuen, Taiwa-cho, Kurokawa-gun, 981-3298 Miyagi, Japan

ARTICLE INFO

Article history:
Available online 7 May 2010

Keywords:
Dynamical systems
Inter-modal mapping
Recurrent neural network with parametric bias
Generalization

ABSTRACT

A system for mapping between different sensory modalities was developed for a robot system to enable it to generate motions expressing auditory signals and sounds generated by object movement. A recurrent neural network model with parametric bias, which has good generalization ability, is used as a learning model. Since the correspondences between auditory signals and visual signals are too numerous to memorize, the ability to generalize is indispensable. This system was implemented in the "Keepon" robot, and the robot was shown horizontal reciprocating or rotating motions with the sound of friction and falling or overturning motion with the sound of collision by manipulating a box object. Keepon behaved appropriately not only from learned events but also from unknown events and generated various sounds in accordance with observed motions.

1. Introduction

Various kinds of robot systems that interact with humans have recently received a great deal of attention, represented by increased interest in humanoid robots (http://www.honda.co.jp/ASIMO/; Ishiguro et al., 2001), particularly human assistance robots. These robots have to react to multi-modal sensory inputs in order to execute tasks and communicate with human operators. Most humanoid robots developed so far handle the sensory data from different modes independently. After information processing for each modality, the results are synchronized and integrated, a process that is quite difficult to design. An alternative approach is for the robot to handle all the data simultaneously, which is the approach we have taken.

People deal with "cross-modal information" by, for example, expressing auditory information (e.g. sounds of collision) by using visual expressions like gestures (e.g., moving the hand quickly and stopping it sharply). These gestures are apparently related to the development of onomatopoeia (Werner and Kaplan, 1963). We call this process "inter-modality mapping."

Arsenio and Fitzpatrick proposed an interesting method for object recognition using "periodic dynamics" in multi-modal information (Arsenio and Fitzpatrick, 2003). Using this method, a humanoid robot called Cog recognizes objects by coupling data from different modes. For example, a hammer is recognized from its visual image, ringing sound, and hitting motion. The crucial concept of this method regards recognition as the extraction of common dynamics from multi-modal sensory information. However, the targets are restricted to rhythmic patterns.

Our ultimate goal is to design and implement a method for inter-modal mapping. The method should enable a robot to generate motion from various types of sound signals and to generate sound appropriate to various types of images. Such a method should lead to various interesting findings in the field of cognitive sciences.

Section 2 presents our model of inter-modality mapping—a robot acquires the relationships between different items of modal information by observing various events. Section 3 introduces the neural network model used for association/translation between inter-modalities and for the generalization of multi-modal sensory dynamics obtained from observation experience. Section 4 describes the implementation of our system in a small robot called Keepon. Section 5 presents the experimental results for inter-modality mapping, and Section 6 discusses the generalization ability of our method on the basis of the results of experiments using environmental sound. Section 7 summarizes the key points and mentions future work.

2. Model of inter-modality mapping

As mentioned above, conventional robot systems typically process sensor modalities separately. However, various modes of sensory information are usually received simultaneously. We have developed a procedure for interpretation of inter-modality mapping that is divided into two main phases, a "learning phase" and a subsequent "interaction phase."

* Corresponding author. Fax: +81 75 753 5386.
E-mail addresses: ogata@i.kyoto-u.ac.jp (T. Ogata), nishide@i.kyoto-u.ac.jp (S. Nishide), xkozima@myu.ac.jp (H. Kozima), komatani@i.kyoto-u.ac.jp (K. Komatani), okuno@i.kyoto-u.ac.jp (H.G. Okuno).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.05.002

2.1. Learning phase ("looking at sound source")

In the learning phase, the robot observes an event that can have various kinds of sound, such as a bouncing sound, a friction-induced sound, a continuous sound, or a rhythmic sound (see Fig. 1). The robot memorizes these sounds along with the motions of the sound source. We call this the "robot looking at sound source" phase.

2.2. Interactive phase ("mapping from sound to motion")

In the interactive phase, the sensory information from a single modality (image or sound) is input into the robot's system. The robot associates this information with the information from the other modalities and expresses it by, for example, moving its body to create the same motion as the sound source (Fig. 2). Conversely, the robot observes a motion and outputs the sound memorized for that motion (Fig. 3).

3. Recurrent neural network model

3.1. Introduction

There are numerous sounds in the environment around us. It is impossible to construct a database that can systematically store all environmental sounds. Therefore, to achieve inter-modal mapping, the robot must be able to generalize various sounds from a limited collection of memorized sounds. That is, the robot should be able to adapt to unknown stimuli.

To meet this requirement, we use the artificial neural network model proposed by Tani and Ito (2003). The main characteristic of this recurrent neural network model with parametric bias (RNNPB model) is that chunks of sequence patterns of the sensory-motor flow can be represented by a vector of small dimension. This vector plays the same role as bifurcation parameters in nonlinear dynamic systems. That is, different vector values result in different dynamic patterns being generated by the system. The main advantage of using parameter bifurcation is that ideally the RNNPB model can encode an infinite number of dynamic patterns with modulated analog values of the vector.

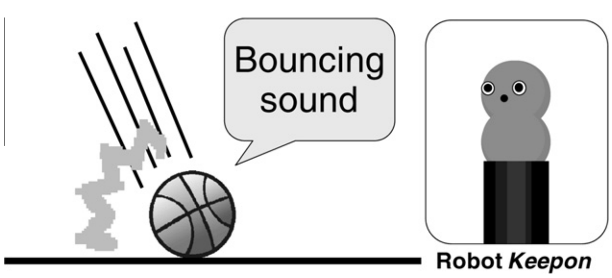

Fig. 1. Learning phase: robot looking at sound source.

An RNNPB model is usually designed as a predictor ("forwarding-forward model") for which input is current condition $S(t)$ and output is next condition $S(t+1)$. Its network has the same structure as the Jordan-type RNN (Jordan, 1986) except that it has parametric bias (PB) nodes in the input layer (see Fig. 4). Unlike other input nodes, these PB nodes have a constant value throughout each time sequence. The context layer has a loop that inputs the current output as input data into the next step. This enables the RNNPB model to learn the time sequences on the basis of past contexts.

The RNNPB model has three activation modes: learning, recognition, and generation based on prediction. Learning is a process that modulates the network weights and PB values by using output error. Recognition is a process that inputs the whole sequence to output the PB representing the sequence. Prediction is a process that inputs a certain state of a given sequence to output the next state.

3.2. Learning mode

In the learning mode, it updates its weights and the value for the "parametric bias' simultaneously using the back-propagation through time (BPTT) method (Rumelhart et al., 1986) with prediction error. Each update is carried out using the equations below. The step length of a sequence is denoted by l. For each sensory-motor output, back-propagated errors with respect to PB nodes are accumulated and used to update the PB values. The update equations for the ith unit of the parametric bias at the t in sequence are

$$\delta\rho_t = k_{bp} \cdot \sum_{t-l/2}^{t+l/2} \delta_t^{bp} + K_{nb}(\rho_{t+1} - 2\rho_t + \rho_{t-1}), \tag{1}$$

$$\rho_{t+1} = \rho_t + \varepsilon \cdot \delta\rho_t, \tag{2}$$

$$p_t = sigmoid(\rho_t/\zeta). \tag{3}$$

In Eq. (1), the $\delta\rho_t$ for updating the internal values ρ_t of the PB (p_t) is obtained from the summation of two terms. The first term represents the delta error, δ_t^{bp}, back-propagated from the output nodes to the PB nodes: it is integrated over the period from the $t - l/2$ to the $t + l/2$ steps. Integrating the delta error prevents local fluctuations in the output errors from significantly affecting the temporal PB values. The second term is a low-pass filter that inhibits frequent rapid changes in the PB values. Internal value ρ_t is updated using the $\delta\rho_t$, as shown in Eq. (2). The k_{bp} (>0), k_{nb} (<0), and ε (>0) are coefficients. The current PB values are obtained from the sigmoidal outputs of the internal values. After learning the time sequences, the RNNPB model self-organizes the PB values at which the specific properties of each individual time sequence are encoded and can generate a sequence from the corresponding PB values. Our goal is to identify the specific parameter values corresponding to each event. Therefore, to fix the parameter values during the motion recognition, parameter k_{nb} in Eq. (1) was set to 0 in our RNNPB model training:

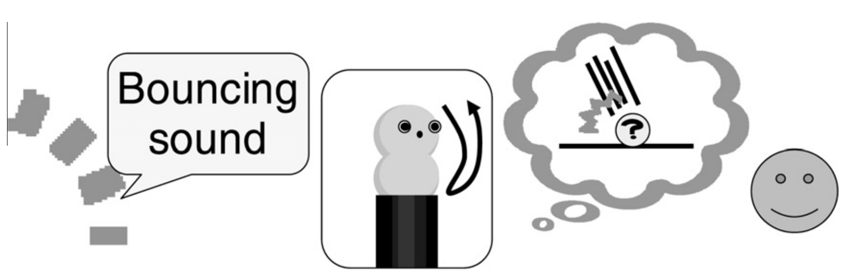

Fig. 2. Interactive phase: mapping from sound to motion.