Journal of Computer and Communications, 2015, 3, 46-51
Published Online November 2015 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2015.311008

An Acoustic Events Recognition for Robotic Systems Based on a Deep Learning Method

Tadaaki Niwa[1*], Takashi Kawakami[2], Ryosuke Ooe[2], Tamotsu Mitamura[3], Masahiro Kinoshita[3], Masaaki Wajima[2]

[1]Graduate School of Engineering, Hokkaido University of Science, Sapporo, Japan
[2]Faculty of Engineering, Hokkaido University of Science, Sapporo, Japan
[3]Faculty of Future Design, Hokkaido University of Science, Sapporo, Japan
Email: *r13301@hus.ac.jp

Received August 2015

Abstract

In this paper, we provide a new approach to classify and recognize the acoustic events for multiple autonomous robots systems based on the deep learning mechanisms. For disaster response robotic systems, recognizing certain acoustic events in the noisy environment is very effective to perform a given operation. As a new approach, trained deep learning networks which are constructed by RBMs, classify the acoustic events from input waveform signals. From the experimental results, usefulness of our approach is discussed and verified.

Keywords

Acoustic Events Recognition, Deep Learning, Restricted Boltzmann Machine

1. Introduction

Social insect, or social animal can work more than their own ability in concert with other individuals. They usually communicate with each other through sounds and vibrations. Also, certain animals recognize events of surrounding environment using acoustic information that is obtained from the environmental sounds.

The other hand, multiple autonomous robots systems or swarm robots systems are needed to develop for the disaster response and search-rescue missions. We know well that the terrible disaster of nuclear plant in Japan reminding of necessity for robotic response systems. These systems are expected to achieve the difficult missions by cooperating among relatively simple robots. In this case, detecting and recognizing the environmental information is very important functions in whole system. Usually, vision-based recognition mechanisms are adopted in autonomous robotic systems. However, in swarm robots systems, each robot has comparatively simple structure without a camera, and each robot will act on the basis of the locally information to which it can be easily acquired. Also, sound information beyond a wall cannot be recognize by only vision-based systems. For example, it is very effective to detect and recognize the explosion sounds or human voices from the other side of the wall in the noisy environment. Therefore, we focus on developing the classification and recognition me-

*Corresponding author.

How to cite this paper: Niwa, T., Kawakami, T., Ooe, R., Mitamura, T., Kinoshita, M. and Wajima, M. (2015) An Acoustic Events Recognition for Robotic Systems Based on a Deep Learning Method. *Journal of Computer and Communications*, **3**, 46-51. http://dx.doi.org/10.4236/jcc.2015.311008

chanisms of acoustic events.

Recognizing acoustic events are becoming a key component of multimedia computational systems of all types, including robotic systems. Until now, identifying real-world acoustic events are tried by using some methodologies, e.g., a layered Hidden Markov Model (HMM).

In real environments, it is necessary to consider that an observed sound includes multiple sound source and are mixed their sound source. For example, a sound of environment that surrounds a living space is mixed a voice, a music, a engine sound of car, and a other living sound. Therefore, it is important to separate sound source or detect typical sound at a certain timing. In this paper, we focused on detection of typical sound at a certain timing

In this paper, we discuss the acoustic events classification and recognition mechanisms based on the deep learning structure. This structure is constructed by Restricted Boltzmann Machines (RBM). As the experiments, we configured a deep network based on convolutional RBM and convolutional Deep Belief Nets. Learning and classifying results of model are compared, and discussed.

2. Restricted Boltzmann Machine

2.1. Binary Visible Units and Binary Hidden Units

An RBM [1] [2] is an undirected graphical model that is used to describe the dependency among a set of random variables over a set of observed data. In this model, the stochastic visible units $\boldsymbol{v}$ connected to the stochastic hidden units h. The joint distribution $p(v, h)$ over the visible units and hidden units is defined through energy function $E(v, h)$:

$$p(v,h) = \frac{1}{Z} e^{-E(v,h)} \tag{1}$$

and the probability density $p(v)$ over the visible units defined as:

$$p(v) = \frac{1}{Z} \sum_{h} e^{-E(v,h)} \tag{2}$$

$$Z = \sum_{v,h} e^{-E(v,h)} \tag{3}$$

where Z is normalization factor (or partition function) that can be estimated by the annealed importance sampling (AIS) method.

The commonly case (where $v = [v_1, v_2, ..., v_i], v_I \in \{0,1\}$ and $h = [h_1, h_2, ..., h_j], h_J \in \{0,1\}$), the energy function $E(v, h)$ of an RBM is defined as:

$$E(v,h) = -\sum_{i} b_i v_i - \sum_{j} c_j h_j - \sum_{i} \sum_{j} v_i W_{ij} h_j \tag{4}$$

where v_i, h_j are states of visible unit i and hidden unit j, b_i, c_j are their biases and W_{ij} is the weight between them. Since, an RBM has no intra-layer connections, the visible unit activations and the hidden unit activations are mutually conditional independence. Therefore, the conditional probability $p(v_i|h)$ and $p(h_j|v)$ that activate each unit are represented by a simple functions as:

$$p(v_i \mid h) = sigmoid(b_i + \sum_{j} h_j W_{ij}) \tag{5}$$

$$p(h_j \mid v) = sigmoid(c_j + \sum_{i} v_i W_{ij}) \tag{6}$$

where $sigmoid(x) = 1/(1 + e^{-x})$ is standard sigmoid function.

2.2. Gaussian Visible Units

For real-valued data such as natural images or the Mel-Frequency Cepstral Coefficients, Bernoulli-Bernoulli (or binary-binary) form is poor representation. However, RBM can be applied to model the distribution of real-valued data by adopting its Gaussian-Bernoulli (or Gaussian-binary) form [3] [4]. Where $v = [v_1, v_2, ..., v_I], v_i \in R$ and

$h = [h_1, h_2, ..., h_J], h_i \in \{0,1\}$. In this case, the energy function $E(v, h)$ of an RBM is defined as:

$$E(v,h) = \sum_i \frac{(v_i - b_i)}{2\sigma_i^2} - \sum_j c_j h_j - \sum_i \sum_j \frac{v_i}{\sigma_i} W_{ij} h_j \tag{7}$$

and the conditional probability $p(v_i|h)$ is defined as:

$$p(v_i \mid h) = N(v \mid b_i + \sum_j h_j W_{ij}, \sigma_i^2) \tag{8}$$

where $N(x;\mu;\sigma^2)$ is Gaussian probability density with mean μ and variance σ^2, and σ^2 is variance parameter of Gaussian noise on visible unit *i*.

2.3. Contrastive Divergence Learning Algorithm

The CD-k algorithm [5] is fast calculation algorithm to approximate the gradients of log-likelihood. Given a set of training data, the model parameters $\theta = \{W, b, c\}$ of an RBM are estimated by maximum likelihood learning of $\boldsymbol{p(v)}$. The model parameters that maximize the log-likelihood are determined with stochastic gradient method in general. The gradient of this log-likelihood is given through energy function $E(v, h)$ of an RBM:

$$\frac{\partial \ln L(\theta \mid v)}{\partial \theta} = -\sum_h p(h \mid v) \frac{\partial \ln E(v,h)}{\partial \theta} + \sum_{v,h} p(h,v) \frac{\partial \ln E(v,h)}{\partial \theta} \tag{9}$$

However, this gradient is difficult to calculate strictly, because, calculation cost increase exponentially. CD algorithm approximate the gradients of log-likelihood using k-step Gibbs sampling and joint probability $p(v|h)$, $p(h|v)$. This gradient is given as:

$$\frac{\partial \ln L(\theta \mid v)}{\partial \theta} = -\sum_h p(h \mid v^{(0)}) \frac{\partial \ln E(v,h)}{\partial \theta} + \sum_{v,h} p(h \mid v^{(k)}) \frac{\partial \ln E(v^{(k)},h)}{\partial \theta} \tag{10}$$

Therefore, gradients of each parameter are given as:

$$\frac{\partial \ln L(\theta \mid v)}{\partial W_{ij}} = p(h_j \mid v^{(0)}) v_i^{(0)} - p(h_j \mid v^{(k)}) v_i^{(k)} \tag{11}$$

$$\frac{\partial \ln L(\theta \mid v)}{\partial b_i} = v_i^{(0)} - v_i^{(k)} \tag{12}$$

$$\frac{\partial \ln L(\theta \mid v)}{\partial c_j} = p(h_j \mid v^{(0)}) - p(h_j \mid v^{(k)}) \tag{13}$$

3. Experiments and Results

3.1. Experiment Condition

We configured a deep neural network based on convolutional RBM [6] and convolutional Deep Belief Nets (CDBN) [7] as shown **Figure 1**. Network has three convolution layers, two max pooling layers, and one full connection layer. Each layer setting illustrate **Table 1**. We configured a Convolutional Neural Network (CNN) [8] [9] of same parameters for comparison.

In pre-training step, each layer is training as standard RBM using the patch that was cut out from the inputs. Because, to reduce the computational cost. CD learning with 1-step Gibbs sampling (CD1) was adopted for the RBM training and the learning rate was 0.0001. The batch size was set to 100 and 100 epochs were executed for estimating each RBM. In fine-tuning step and training CNN, we used Adam learning method [10] and early stopping.

We used train and test dataset of D-CASE challenge [11] for our experiments. This data set is recorded typical 16 category sounds of office environments. Also, training data and test data has been granted noise. 256-order spectrograms were derived from the waveform by STFT analysis using 512 points hamming window at 10 milliseconds frame shift. We also constructed 256 × 100 milliseconds-order patches of spectrogram from spectrogram using 50 milliseconds frame shift.

3.2. Results and Discussions

In the experiments, the network has over-fitting **(Figure 2** and **Figure 3**). Also, transition of each values are analogous at between fine tuning of CDBN and CNN. In the classification of results after learning, f-measure of each network is not the significance compared to the case of random (**Figure 4**). We believe that there is cause to representation of the input data by transition of each value in training data and the outcome of deep learning

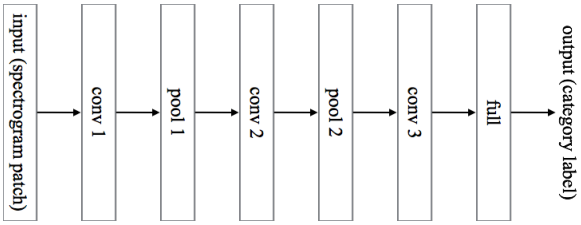

Figure 1. Structure of network for our experiments.

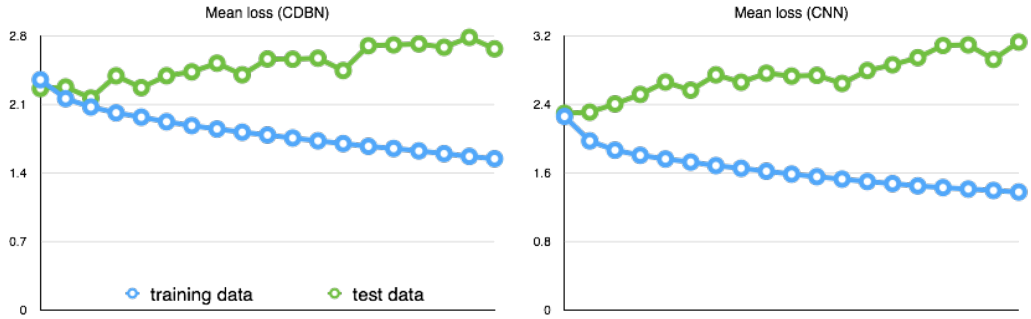

Figure 2. Learning curve (mean loss of cross entropy).

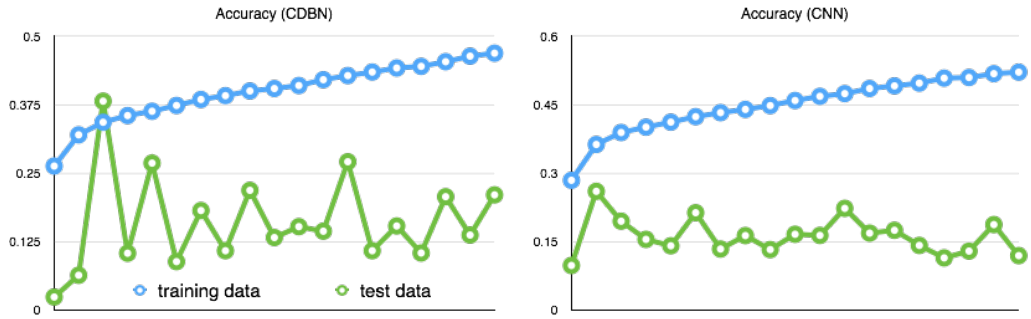

Figure 3. Learning curve (accuracy).

Table 1. Setting of each layers.

Layer names	Layer parameters			
	Filter size (w × h)	Output map size	Stride	Function
data	-	257 × 10 × 1	-	
conv 1	5 × 2	253 × 9 × 16	1 × 1	
pool 1	2 × 2	127 × 5 × 16	2 × 2	ReLU
conv 2	5 × 2	123 × 4 × 16	1 × 1	-
pool 2	2 × 2	62 × 2 × 16	2 × 2	ReLU
conv 3	5 × 2	58 × 1 × 16	1 × 1	-
full 1	-	1 × 1 × 17	-	Soft-max

conv, pool and full are convolution layer, pooling layer and full connection layer respectively.

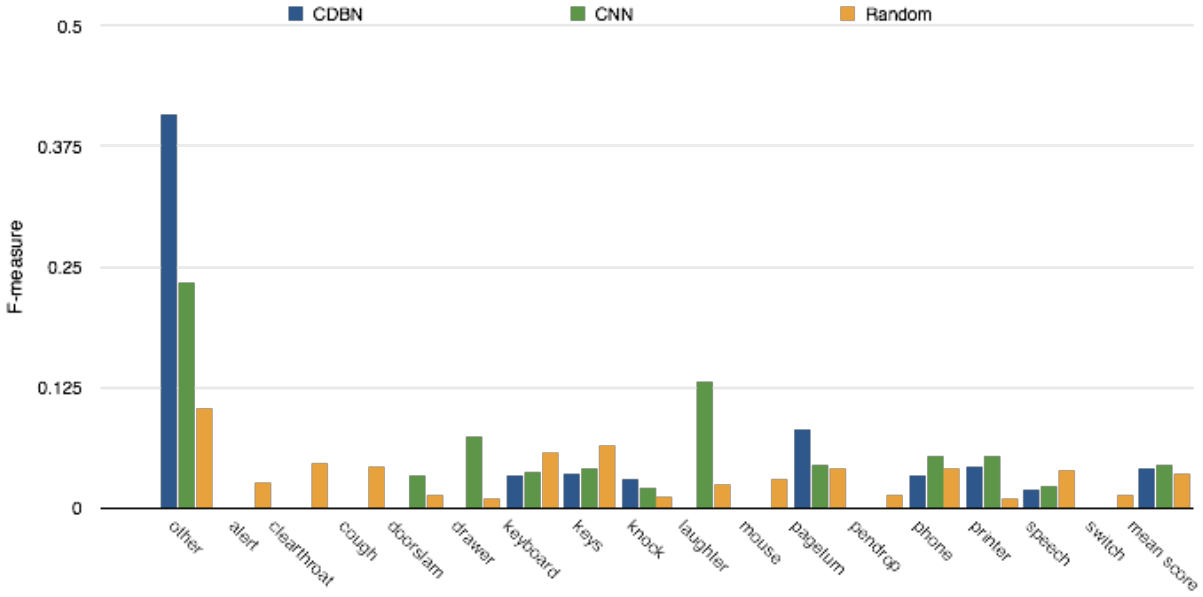

Figure 4. F-measure score of each category.

in the recent years. In recent years, it is known that it is possible to extract better features from the spectrogram of FFT by deep learning. However, in all cases, it is suggested that it may not be effective. If the frequency analysis by the auditory filter in consideration of the hearing mechanism of organisms, we think that the better results are obtained.

4. Conclusions

In swarm robots systems, each robot has comparatively simple structure without a camera, and each robot will act on the basis of the locally information to which it can be easily acquired. In this case, detecting and recognizing the environmental information is very important functions in whole system. Therefore, we focus on developing the classification and recognition mechanisms of acoustic events for swarm robots.

In this paper, we proposed the acoustic events classification and recognition mechanisms based on the deep learning structure. This structure is constructed based on RBM. However, in the experiments on this paper, we cannot figure out how to get enough recognition accuracy in noise environments. We believe that there is cause to representation of the input data by transition of each value in training data and the outcome of deep learning in the recent years. If the frequency analysis by the auditory filters in consideration of the hearing mechanism of organisms, we think that the better results are obtained.

In the future work, we plan to incorporate the auditory filter to our approach, and will expect to improve the recognition accuracy by this plan.

References

[1] Hinton, G.E., Osindero, S. and Teh, Y.W. (2006) A Fast Learning Algorithm for Deep Belief Nets. *Neural computation*, **18**, 1527-1554. http://dx.doi.org/10.1162/neco.2006.18.7.1527

[2] Freund, Y. and Haussler, D. (1994) Unsupervised Learning of Distributions of Binary Vectors Using Two Layer Networks. *Computer Research Laboratory* [University of California, Santa Cruz].

[3] Hinton, G.E. and Salakhutdinov, R.R. (2006) Reducing the Dimensionality of Data with Neural Networks. *Science*, **313**, 504-507. http://dx.doi.org/10.1126/science.1127647

[4] Cho, K., Ilin, A. and Raiko, T. (2011) Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines. In: *Artificial Neural Networks and Machine Learning—ICANN* 2011, Springer Berlin Heidelberg, 10-17. http://dx.doi.org/10.1007/978-3-642-21735-7_2

[5] Hinton, G.E. (2002) Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, **14**, 1771-1800. http://dx.doi.org/10.1162/089976602760128018

[6] Norouzi, M., Ranjbar, M. and Mori, G. (2009) Stacks of Convolutional Restricted Boltzmann Machines for Shift-Invariant Feature Learning. *IEEE Conference on Computer Vision and Pattern Recognition*, 2735-2742.

[7] Lee, H., Grosse, R., Ranganath, R. and Ng, A.Y. (2009) Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations. *Proceedings of the* 26*th Annual International Conference on Machine Learning*, 609-616. http://dx.doi.org/10.1145/1553374.1553453

[8] Simard, P.Y., Steinkraus, D. and Platt, J.C. (2003) Best Practices for Convolutional Neural Networks Applied to Visual Document Analysis. In: *null*, 958. http://dx.doi.org/10.1109/icdar.2003.1227801

[9] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 1097-1105.

[10] Kingma, D. and Ba, J. (2015) Adam: A Method for Stochastic Optimization. *Proceedings of the* 3*rd International Conference on Learning Representations* (*ICLR* 2015).

[11] IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events. http://c4dm.eecs.qmul.ac.uk/sceneseventschallenge/