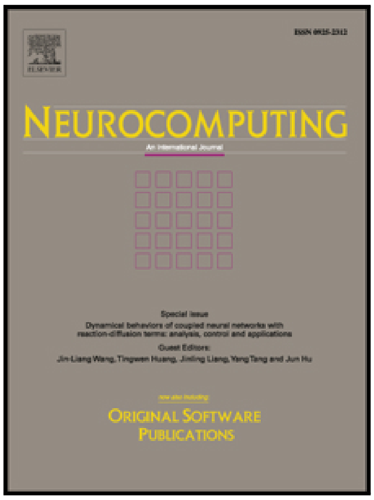
Accepted Manuscript



Knock-Knock: Acoustic Object Recognition by using Stacked Denoising Autoencoders

Shan Luo , Leqi Zhu , Kaspar Althoefer , Hongbin Liu

PII: S0925-2312(17)30509-X
DOI: 10.1016/j.neucom.2017.03.014
Reference: NEUCOM 18235

To appear in: *Neurocomputing*

Received date: 9 October 2016
Revised date: 26 January 2017
Accepted date: 6 March 2017

Please cite this article as: Shan Luo , Leqi Zhu , Kaspar Althoefer , Hongbin Liu , Knock-Knock: Acoustic Object Recognition by using Stacked Denoising Autoencoders, *Neurocomputing* (2017), doi: 10.1016/j.neucom.2017.03.014

Knock-Knock: Acoustic Object Recognition by using Stacked Denoising Autoencoders

Shan Luo, Leqi Zhu, Kaspar Althoefer, Hongbin Liu

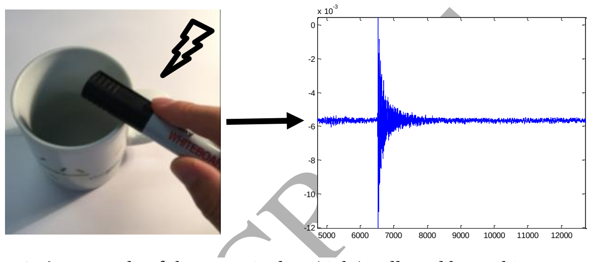


Fig. 1. An example of the acoustic data (right) collected by striking a marker pen on a cup (left).

***Abstract*—This paper presents a successful application of deep learning for object recognition based on acoustic data. It can restrict capability of the representation to serve different applications and may only capture insignificant characteristics for a task when using handcrafted features. In contrast, there is no need to define the feature representation format when using multilayer/deep architecture methods and features can be learned from raw sensor data without defining discriminative characteristics a-priori. In this paper, stacked denoising autoencoders are applied to train a deep learning model. Thirty different objects were classified in our experiment and each object was knocked 120 times by a marker pen to obtain the auditory data. By employing the proposed deep learning framework, a high accuracy of 91.50% was achieved. The traditional method using handcrafted features with a shallow classifier was taken as a benchmark and the attained recognition rate was only 58.22%. Interestingly, a recognition rate of 82.00% was achieved when using a shallow classifier with raw acoustic data as input. Nevertheless, the time taken for classifying one object using deep learning was far less (6.77 faster) than utilizing this method. It was also explored how different model parameters in deep architecture would affect the recognition performance.**

***Keywords* — Object recognition, deep networks, acoustic data analysis.**

I. INTRODUCTION

Future intelligent robots are envisioned to be endowed with perceptive capabilities to see, touch and hear what is happening in the ambient world. It enables robots perform various tasks and object recognition is among the most common and significant ones. To perform this task, many types of sensors can be utilized and each kind of sensor offers a different view of objects. One of the richest and most widely used sensors is the camera as much information can be acquired from one single image. Because of this, vision has attracted considerable attention in object recognition by classifying the color [1], texture [2], [3], surface reflectance [4] and appearance [5], [6]. But vision is heavily dependent on the surrounding environments and would fail due to the variance of poses, illumination changes or occlusion by other objects. Another not as powerful but also very rich sensing modality is the sense of touch. With the use of force/tactile sensors, the object properties can be revealed by accessing the information of hardness/softness [7], elasticity [8], thermal cues [9], surface texture [10] and surface friction properties [11]. However, the tactile object recognition needs the direct contact with the objects and there will be a risk of damage when handling objects of hazardous materials. As an alternative acoustic data can be acquired by sensors like microphones to recognize the objects. It can allow the robot to work in safer conditions. In addition, sound signals generated by striking an object can expose the intrinsic properties of objects such as elasticity and internal friction [12]. The elasticity of an object is directly related to the speed of sound waves in the object and therefore influences the frequency of the sound. The internal friction, or dampness, determines how the generated sounds decay over time [12] and provides shape-variant acoustic features for object classification [13].

To date, however, audition has been largely neglected relative to vision and tactile sensation in the application of object recognition. One of the most dominant factors is that the auditory data is more abstract compared to visual images and force/tactile data. In the traditional acoustic based recognition, the task is achieved by using handcrafted features in time [14] or frequency domain [15] with shallow classifiers. However, there are several drawbacks of methods in this manner. Firstly, it is time consuming and laborious to extract the features. Secondly, it is difficult to design appropriate features oriented by different tasks. Thirdly, using features of pre-defined types can restrict capability of the representation to serve different applications and may result in capturing characteristics of minor importance for a task. Fourthly, as for vision and tactile sensing, acoustic features are present in a hierarchical structure, therefore, the use of handcrafted features and shallow classifiers will cause information

Shan Luo (corresponding author), Leqi Zhu, Kaspar Althoefer, Hongbin Liu are with the Centre for Robotics Research, Department of Informatics, King's College London, WC2R 2LS, UK (e-mail: shan.luo@kcl.ac.uk, leqi.zhu@kcl.ac.uk, k.althoefer@kcl.ac.uk, hongbin.liu@kcl.ac.uk).