Journal of Fiber Bioengineering and Informatics 7:4 (2014) 603–613
doi:10.3993/jfbi12201413

Learning Block Group Sparse Representation Combined with Convolutional Neural Networks for RGB-D Object Recognition ⋆

Shuqin Tu, Yueju Xue*, Jinfeng Wang, Xiaolin Huang, Xiao Zhang

College of Information, South China Agricultural University, Guangzhou 510642, China

Received 5 June 2014; accepted (in revised version) 30 Nov 2014; available online 17 December 2014

Abstract

RGB-D (Red, Green and Blue-Depth) cameras are novel sensing systems that can improve image recognition by providing high quality color and depth information in computer vision. In this paper we propose a model to study feature representation of combined Convolutional Neural Networks (CNN) and Block Group Sparse Coding (BGSC). Firstly, CNN is used to extract low-level features from raw RGB-D images directly by applying unsupervised algorithm. Then, BGSC is used to obtain higher feature representation for classification by incorporating both the group structure for low-level features and the block structure for the dictionary in subsequent learning processes. Experimental results show that the CNN-BGSC approach has higher accuracy on a household RGB-D object dataset by linear predictive classifier than using Convolutional and Recursive Neural Networks (CNN-RNN), Group Sparse Coding (GSC), and Sparse Representation base Classification (SRC).

Keywords: RGB-D; Convolutional Neural Networks; Block Group Sparse Coding; Classification Recognition; Feature Learning Methods

1 Introduction

Object recognition has attracted numerous cross-field attentions in computer science, appealing to researchers from fields such as computer vision, machine learning and robotics. In the past few decades, a variety of features and classification algorithms have been proposed and applied to improve the technique, resulting in significant progress in object recognition capabilities, as are evident from steady improvements on standard benchmarks such as Caltech101 [1] and CIFAR [2]. New sensing technologies, the rapidly maturing technologies of RGB-D (Kinect-style) and depth cameras [3, 4] provide synchronized videos of color and depth with high quality, presenting

⋆Project supported by National Science and Technology Support Program of China (2013BAJ13B05) the Joint Funds of the National Natural Science Foundation of China under Grant No. U1301253.

*Corresponding author.
Email address: xueyueju@163.com (Yueju Xue).

1940–8676 / Copyright © 2014 Binary Information Press & Textile Bioengineering and Informatics Society
December 2014

a great opportunity for combining color and depth based on recognition. Most recent methods for object recognition with RGB-D images use hand-designed features which include SIFT for two dimensional (2D) images [5], Spin Images [6] for three dimensional (3D) point clouds, specific color, shape and geometry features [7, 8], and a Learned Feature Descriptor [9, 11].

Although feature learning methods listed above have achieved some success, they are still rather slow and require additional input channels such as surface normal or texture feature and carefully design specific feature learning method for each specific task. To make features automatically fit to object recognition task, researchers employed deep learning framework to achieve feature representation from raw data. Convolutional and Recursive Neural Networks (CNN-RNN) model was proposed [10], obtaining good performance for classifying RGB-D images. However, the main shortcoming of CNN-RNN model is that the extracted features are still not suitable for classification directly due to large dimensions.

Features captured from images are actually sparse [24]. The classifier for object recognition only needs to extract features which can capture the essence of an image by removing irrelevant information and noises. Therefore the problem in representation learning lies in sparse coding. Furthermore, these features may group cluster in which the nonzero elements exist in the union of subspaces. The group sparse coding in which a block structure is imposed on the dictionary [12] is widely used in image classification [13, 14]. we introduced Convolutional Neural Networks (CNN) [10] for extracting lower features from raw RGB-D images and structured group sparse representation for learning higher representation.

In this paper, we propose an efficient model based on a combination of Convolutional Neural Networks (CNN) and Block Group Sparse Representation (BGSC) called CNN-BGSC for feature extraction. The CNN can extract low-level invariant features from raw RGB-D images; these features are then given as inputs to BGSC to compose higher features. The objective of this study is to explore higher feature representation by combining the CNN and BGSC methods for RGB-D object recognition.

This paper is organized as follows. A brief introduction and related work are presented in Section 1 and Section 2, respectively. The method by combining CNN model and structured block/group sparse coding will be presented in Section 3. Then the experiment and results will be shown in Section 4. Section 5 is the conclusion.

2 Relate Work

We will briefly review current approaches for recognition on RGB-D data as well as the current state of feature extraction from images. The most common RGB-D recognition methods are based on extracted orientation or histograms features such as SIFT [15], SURF [16], textons and SVM classifier. Despite their commonality, it is difficult to design an effective classifier to incorporate additional information. There have been some attempts to overcome the shortcoming by extending SIFT to the depth channel [2]. A recently proposed kernel descriptor [9] provides a general design pattern for combining different features descriptor such as 3D shape, physical size of the object, depth edges, gradients, etc. In [5], Lai et al. proposed a view-based sparse distance learning mechanism. This method can select representative views as object models by applying group lasso regularization. They combined a number of features from 2D and 3D object classification into one descriptor.