Information Sciences 320 (2015) 333–345

Contents lists available at ScienceDirect

Information Sciences

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

Recognizing multi-view objects with occlusions using a deep architecture

Yingjie Xia [a,b], Luming Zhang [c,*], Weiwei Xu [b], Zhenyu Shan [b], Yuncai Liu [b]

[a] College of Computer Science, Zhejiang University, 310027 Zhejiang, China
[b] Hangzhou Institute of Service Engineering, Hangzhou Normal University, 310012 Zhejiang, China
[c] School of Computer and Information, Hefei University of Technology, China

ARTICLE INFO

Article history:
Received 15 June 2014
Received in revised form 18 January 2015
Accepted 30 January 2015
Available online 11 February 2015

Keywords:
Deep learning
Depth data
Multi-view
Object recognition
RGB-D

ABSTRACT

Image-based object recognition is employed widely in many computer vision applications such as image semantic annotation and object location. However, traditional object recognition algorithms based on the 2D features of RGB data have difficulty when objects overlap and image occlusion occurs. At present, RGB-D cameras are being used more widely and the RGB-D depth data can provide auxiliary information to address these challenges. In this study, we propose a deep learning approach for the efficient recognition of 3D objects with occlusion. First, this approach constructs a multi-view shape model based on 3D objects by using an encode–decode deep learning network to represent the features. Next, 3D object recognition in indoor scenes is performed using random forests. The application of deep learning to RGB-D data is beneficial for recovering missing information due to image occlusion. Our experimental results demonstrate that this approach can significantly improve the efficiency of feature representation and the performance of object recognition with occlusion.

1. Introduction

In computer vision applications, object recognition based on multi-view images is essential for extracting semantic information from image pixels. In general, the approaches employed can be categorized into region- and feature point-based recognition methods. In this study, we focus on RGB-D images of indoor scenes where the key task is the recognition of various objects such as desks and chairs. The semantic understanding of RGB-D images can facilitate the 3D modeling of indoor scenes for applications in computer graphics and robotics.

Using RGB-D images, we can extract the color and depth information for an object. Recently, depth information was applied to indoor scene object recognition in [29] using random forests regression. However, this method was only validated with a small-scale database and it was not robust to occlusions in the captured image. Occlusion is one of the major factors that affect the accuracy of object recognition. First, occlusions are difficult to predict during the training stage. Occlusions are determined by the spatial relationships among objects from certain viewpoints, but it is impractical to include all possible occlusions during the training phase of recognition algorithms. In addition, occlusion can significantly decrease the accuracy of the extracted feature vectors because some regions will have missing information.

[*] Corresponding author.
E-mail address: zglumg@gmail.com (L. Zhang).

http://dx.doi.org/10.1016/j.ins.2015.01.038
0020-0255/© 2015 Elsevier Inc. All rights reserved.

In this study, we propose the construction of multi-view shape models of 3D objects with occlusion using an encode–decode deep learning network. Based on these shape models, we can reconstruct the information embedded in an occluded region for use in the subsequent feature representation process. This approach is based on observations that encode–decode deep learning networks are efficient in recovering missing information from partially observed data. Next, we apply random forests to the models for object recognition.

We tested our pipeline using various occluded indoor scene objects with 10–20% occlusion in the object area. Our experimental results demonstrate that preprocessing input depth data using a deep learning network can improve the recognition performance.

The remainder of this paper is organized as follows. In Section 2, we review related research into object recognition and deep learning. Section 3 introduces our approach to 3D object depth data generation, including the methods for training set generation and test set generation. Section 4 describes feature representation using deep learning and in Section 5, we explain how these features are used for 3D object recognition. In Section 6, we present the results of experiments that demonstrate the accuracy of object recognition with and without occlusions. Finally, we give our conclusions and suggestions for future research in Section 7.

2. Related work

2.1. Object recognition

The general process employed for object recognition can be categorized into five phases: verification, detection and localization, classification, naming, and description [18]. Many previous studies have attempted to address the object recognition problem.

Zerroug and Nevatia [40] proposed a generalized cylinder-based recognition method that uses a subset of generalized cylinders to detect the target object [3]. Brooks [4] proposed a parts-based recognition system for object recognition.

Peng et al. [25] proposed a descriptor that utilizes sectors and contour edges to represent local image features. Pham [26] proposed a recognition algorithm based on template matching strategy, which greatly improved the recognition accuracy in chickens. Lu et al. [21] developed a 3D-model based retrieval and recognition method, which uses a semi-supervised learning method to train a classifier for object recognition.

Around 1990, geometric invariants were introduced to build an efficient indexing mechanism for use in recognition. The geometric invariants-based method was first proposed by Schwartz and Sharir [28], where the basic idea is to obtain a coordinates frame using feature points before utilizing the coordinate frame in the expression of an affine invariant. To integrate this approach with 3D modeling, Flynn and Jain [7] proposed the application of invariant feature indexing in 3D object matching.

Subsequently, researchers showed that it is beneficial to employ local features extracted from images for representation during object recognition. An example is the iconic representation algorithm [27], which extracts local feature vectors to represent the multi-scale local orientation of image locations. The best-known method is SIFT [37], which detects the points of interest in an image. Mikolajczyk and Schmid [22] proposed an affine-invariant interest point detector to improve the reliability of recognition using feature grouping modules. Later, Csurka et al. [5] and Sivic and Zisserman [31] introduced the bag-of-features approach for recognition, which obtains satisfactory results with viewpoint changes and background clutter.

2.2. Deep learning

In machine learning, a single architecture such as kernel machines can only work with a fixed feature layer. At present, deep learning is becoming more prevalent, which is a complex architecture with multiple layers that contain nonlinear components with many trainable parameters, [2]. Convolutional neural networks (CNNs) [16] and deep belief networks (DBNs) [12] are two approaches in this area.

In early trials, CNNs proved to be the most successful in applying multi-layer neural networks. These methods use local connections and shared weights to combine multiple neural units [16]. By reducing the number of training parameters, CNNs can accelerate the learning process during general feed-forward back-propagation training [1].

DBN-based methods were proposed in 2006. DBNs comprise an undirected graph model and a directed graph model, where the former is an associative memory that contains top level units and one level of hidden units, and the latter is a stacked restricted Boltzmann machine (RBM) that contains the remaining layers.

There are many variations of deep learning methods, which are similar to CNNs and DBNs. For example, Ji et al. [14] proposed 3-D CNNs, which use temporal features during network modeling. A convolutional DBN was also proposed by [17].

Deep learning is a generic machine learning tool, which can be applied in many areas. For example, during visual document analysis, MNIST can recognize images using CNNs [30]. In face recognition, a deep learning network can learn high-level facial features to narrow the semantic gap between the low level features. Karnowski et al. [15] utilized a deep spatio-temporal inference network during image classification. In general, 3D CNNs are one of the best performing methods and deep learning-based methods are quite useful for solving recognition problems.