Neurocomputing 000 (2017) 1–7

Contents lists available at ScienceDirect

Neurocomputing

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

Long-range terrain perception using convolutional neural networks

Wei Zhang, Qi Chen*, Weidong Zhang, Xuanyu He

School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China

ARTICLE INFO

Article history:
Received 25 January 2017
Revised 13 July 2017
Accepted 6 September 2017
Available online xxx

Communicated by Wei Wu

Keywords:
Terrain perception
Disparity information
Convolutional neural networks
Robot navigation

ABSTRACT

Autonomous robot navigation in wild environments is still an open problem and relies heavily on accurate terrain perception. Traditional machine learning techniques have achieved good performance for terrain perception; however, most of them require manually designed classifiers, meaning they have a poor generalization ability for learning new unknown environments. In this work, we integrate a deep convolutional neural network (CNN) model with a near-to-far learning strategy to improve the accuracy of terrain segmentation and make it more robust against wild environments. The proposed deep CNN model consists of an encoder and a decoder, which perform downsampling and upsampling for terrain feature extraction, respectively. The near-field terrain information obtained directly from the stereo disparity maps is fed into the CNNs as reference to aid in learning the far-field terrain information. Experimental results on a benchmark dataset demonstrate the effectiveness of the proposed terrain perception method.

1. Introduction

Terrain segmentation refers to the process of dividing scenes in the wild into various regions, such as traversable roads, obstacles, and other ambiguous regions. Terrain segmentation can help autonomous robots to perceive the surrounding topographic conditions and perform path planning toward a goal while avoiding obstacles. Although relevant algorithms such as [1,2] exist, it remains a considerable challenge to segment unknown environments accurately due to the difficulty in perceiving and presenting variations in the environments.

Significant work based on image processing and machine learning methods has been devoted to the problem of terrain perception. Halatci et al. [3] presented a system of multi-sensor terrain classification that trained two “low level” classifiers offline for color, texture, and range features based on maximum likelihood estimation (MLE) and a support vector machine (SVM). Their system achieved accurate terrain classification through classifier fusion of visual and tactile features. Anguelov et al. [4] trained a model from a set of labeled scans offline based on Markov random fields (MRFs), and performed graph-cut inference on the trained MRFs to segment new scenes efficiently. Bradley et al. [5] trained a random forest classifier offline using voxel features (scan-line features, point cloud features, and color features) for terrain classification and ground surface height estimation. All of these algorithms rely on offline training methods and could achieve good performance when the testing scenes were similar to the training scenes; however, they may not perform well in a wild environment. Therefore, Procopio et al. [6] trained a model using online learning methods, based on a near-to-far learning strategy, to improve the generalization capability of their model. The stereo labels and color histogram features were extracted in the near field to train a logistic regression classifier, which then evaluated the remainder of the image to arrive at final terrain predictions.

Early approaches relying on low-level vision cues [7] and manually designed classifiers are being replaced by popular deep learning algorithms. Particularly, with the popularization of convolutional neural networks (CNNs), the representation power of CNNs has led to successful applications in handwritten digit recognition, and speech and image recognition [8–11]. In the DARPA’s Learning Applied to Ground Robots (LAGR) program, the existing pioneering work has attempted to combine a CNN-based classifier with a histogram-based approach to divide scenes into several classes [12–14]. There is now an active interest in semantic pixel-wise labeling [15–19] in which each pixel is labeled with a predefined category. According to the SegNet model proposed by Vijay Badrinarayanan et al. [20], it is known that mapping downsampled feature maps onto images at the original resolution for pixel-wise classification is feasible and can achieve good performance. However, traditional deep models such as SegNet do not work well for terrain perception in the wild because they are trained in a traditional supervised manner offline. This means they cannot generalize well for new, unrecognizable data.

* Corresponding author.
E-mail address: chenqi747@gmail.com (Q. Chen).

https://doi.org/10.1016/j.neucom.2017.09.012
0925-2312/

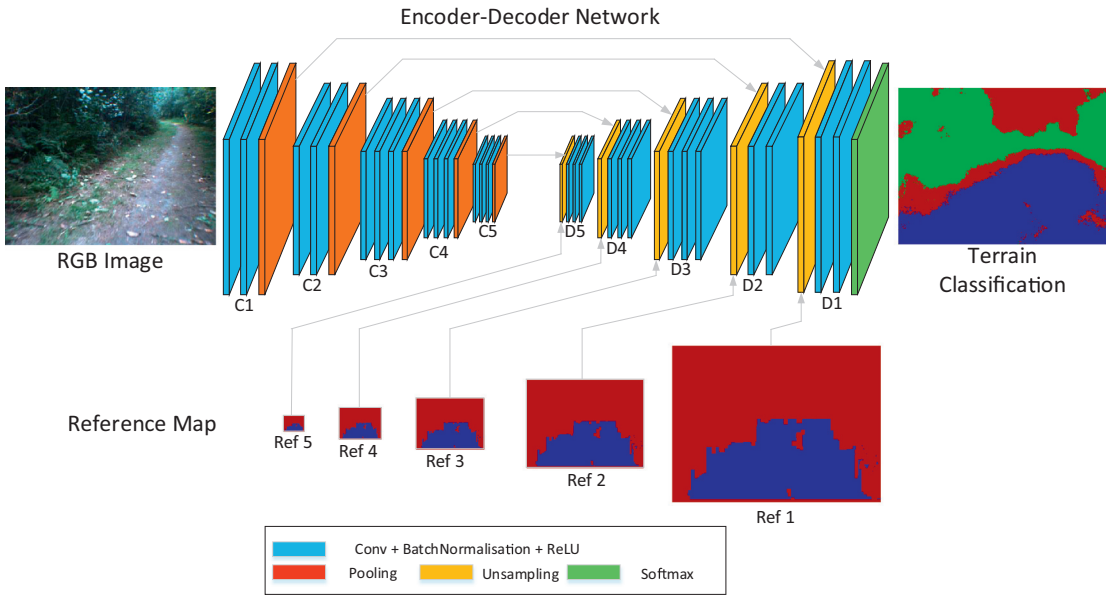

Fig. 1. Proposed deep CNN. The encoder network and decoder network are symmetrical. The decoder upsamples the input feature maps using a deconvolution operation to generate sparse feature maps. The reference map is reduced before being filtered using convolution filters. The feature maps generated from the reference map are concatenated with the sparse feature maps from the decoder. The output of decoder is fed into a softmax classifier for pixel-wise classification.

We introduce a near-to-far strategy combined with CNNs and present an end-to-end training architecture to overcome such limitations. It is demonstrated that training with near-field information strengthens the capabilities of the neural network with respect to learning suitable features for unknown long-field terrain perception. Thus, this work provides an alternative method of using the near-field terrain information for long-range perception. This differs from the traditional work that relied on the near-field information of the current image (one image) to train a classifier online [6] [21]. The potential issue with these methods is that the stereo information is often noisy and sparse, which may make online training unsatisfactory. Additionally, the online training in these methods relies on handcrafted features, which may need parameter tuning in practice. In contrast, the proposed network is trained end to end offline. Thus, it has low computational complexity for testing, and the features are extracted by learning. The network is trained and tested on the LAGR dataset [6] and compared to existing online and offline terrain perception methods.

2. Proposed approach

In this section, we will present the structure of the proposed model, illustrated in Fig. 1, and discuss the effects of reference maps on terrain perception in the wild.

2.1. Architecture

The CNN model consists of an encoder module and a decoder module, followed by a softmax layer. We initialize the encoder parameters using the pretrained weights of the VGG16 network [9]. The output of the decoder module is fed into a multi-class softmax classifier to generate predicted labels for each pixel. The capacity of the decoder module determines the performance of the model. In the decoder network, the input feature maps are upsampled, which produces sparse, expanded feature maps using the memorized max-pooling indices. This upsampling method is called deconvolution.

Similar to SegNet [20], each layer in our encoder module performs convolution using a filter bank and produces a group of feature maps. These feature maps are batch normalized [22], and then

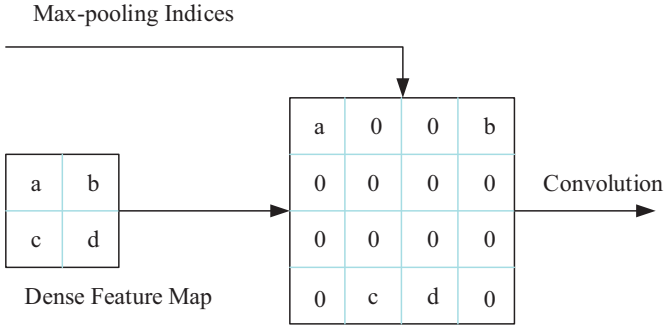

Fig. 2. Illustration of the upsampling process. *a, b, c, d* denote the values of the feature map. This approach uses the max-pooling indices to upsample the feature map(s), which convolves with a trainable filter set. The upsampling process does not involve parameter learning.

a rectified linear unit (ReLU) $max(0, x)$ is applied. Following this convolutional submodule, a max-pooling layer with a 2×2 window and a stride of 2 (nonoverlapping window) is applied, and feature maps are downsampled by a factor of 2. Max pooling is a data compression process that densifies an input representation (images, hidden-layer outputs, etc.) and reduces its dimensionality. Downsampling will produce a large input image context (spatial window) for each pixel in the feature map. The pooling indices are stored for the convenience of upsampling in the decoder module. In the decoder network, the decoder upsamples the input feature maps and produces sparse feature maps using the memorized max-pooling indices. Each feature map in the encoder has a corresponding map in the decoder with the same size. The upsamling process is illustrated in Fig. 2.

We attempt to feed the reference maps, obtained from the disparity maps, into the decoder module to utilize the stereo information. These reference maps are regarded as extra inputs for the model and are convolved with a set of trainable filters. The feature maps generated from the reference maps are concatenated with the sparse feature maps produced by the decoder. Batch normalization (BN) and an ReLU are used following the convolutional layers to suppress overfitting and accelerate the training process. The

Please cite this article as: W. Zhang et al., Long-range terrain perception using convolutional neural networks, Neurocomputing (2017), https://doi.org/10.1016/j.neucom.2017.09.012