Neurocomputing 168 (2015) 454–463

Contents lists available at ScienceDirect

Neurocomputing

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

Deep learning driven blockwise moving object detection with binary scene modeling

Yaqing Zhang, Xi Li*, Zhongfei Zhang, Fei Wu, Liming Zhao

Zhejiang University, China

ARTICLE INFO

Article history:
Received 26 December 2014
Received in revised form
22 May 2015
Accepted 22 May 2015
Communicated by Shuiwang Ji
Available online 2 June 2015

Keywords:
Moving object detection
Deep learning
Feature binarization

ABSTRACT

As an important and challenging topic in computer vision, moving object detection is typically posed as a problem of scene analysis, which usually requires both robust feature description and effective statistical modeling. In general, the conventional detection methods make use of pre-defined hand-crafted features and sophisticated background models for scene analysis. Therefore, they usually have a low generalization capability of adapting to different scenes with diverse spatio-temporal motion information. In the face of high-definition video data, sophisticated statistical modeling often suffers from an expensive computation or memory cost because of its low efficiency in evaluation and parallelization. In order to address this issue, we propose a deep learning based block-wise scene analysis method equipped with a binary spatio-temporal scene model. Based on the stacked denoising autoencoder, the deep learning module of the proposed method aims to learn an effective deep image representation encoding the intrinsic scene information, which leads to the robustness of feature description. Furthermore, the proposed binary scene model captures the spatio-temporal scene distribution information in the Hamming space, which ensures the high efficiency of moving object detection. Experimental results on several datasets demonstrate the effectiveness and efficiency of the proposed method.

1. Introduction

Recent years have witnessed an increasing demand of effectively understanding and compressing high-definition video surveillance data in real-time. Typically, such a task is usually cast as a problem of moving object detection [1], which aims to capture the object motion information in different conditions. Usually, moving object detection is composed of two modules: feature representation and statistical scene modeling. Specifically, the goal of feature representation is to effectively reflect the intrinsic structural properties of scene pixels, while statistical scene modeling aims at building a robust scene model that is capable of adaptively capturing the underlying spatio-temporal distribution information on scene pixels.

In general, the traditional detection methods directly employ a set of hand-crafted features for feature representation, which takes a fixed and pre-defined feature extraction mode [2–8]. For instance, the commonly used color features (e.g., raw pixel intensity [6]) only reflect the visual perception properties of scene pixels, and often ignore the local contextual information, resulting in the sensitivity to noise and illumination changes. The texture features (e.g., LBP [2]) encode the local geometric information on scene pixels, and are only suitable for the particular textural scenes. The gradient-based features (e.g., HoG [5]) are designed to capture the local shape properties of scene pixels, and have a low generalization capability of reflecting the other properties of scene pixels. The transform coefficient features (e.g., 3D Walsh–Hadamard transform [8]) map the scene pixels into a set of orthogonal subspaces, and use the transform coefficients as the feature representation that only encodes some particular transform information on scene pixels. Consequently, the aforementioned features are incapable of well adapting to different types of scenes with diverse appearance characteristics, resulting in a low generalization capacity for scene analysis.

Moreover, most existing methods for statistical scene modeling often take a sophisticated modeling strategy for capturing the spatio-temporal distribution characteristics of scene pixels, and therefore suffer from a high computation or memory cost [6,9–11,10,12–15]. Typically, such modeling methods include mixture of Gaussian model (MOG) [6], kernel density estimation (KDE) model [9], and subspace learning using principal component analysis (PCA) [10]. In theory, these methods are often time-consuming because of their frequent floating operations with high memory usage for high-resolution videos. To make scene modeling more flexible, a number of sparse optimization based scene models [12,13] are recently proposed; but they cannot be highly scalable due to the high complexity induced by the ℓ_1 optimization. In recent

* Corresponding author.
E-mail address: xilizju@zju.edu.cn (X. Li).

http://dx.doi.org/10.1016/j.neucom.2015.05.082
0925-2312/

years, deep learning is applied for dynamic background learning [16]; however, this work does not consider the further transformation for computational efficiency.

In this paper, we propose a deep learning driven method for blockwise moving object detection based on binary scene modeling. The proposed method constructs a deep feature representation by learning a stacked denoising autoencoder. After that, the learned feature representation is mapped into the binary Hamming space, and we construct a density analysis based scoring function for moving object detection. The main contributions of this work are summarized as follows:

1. We introduce deep learning to construct a robust feature representation for moving object detection. The learned feature representation is capable of well capturing the intrinsic structural properties of a scene and adaptively discovering a set of filter patterns that are robust to complicated factors such as noise and illumination variation. Based on the deep learning framework, we further present a hash method to binarize the feature representation for substantially reducing memory usage in the process of dealing with long-duration high-resolution videos. The hash method is very efficient in terms of several simple thresholding operations, which lays a solid foundation for real-time moving object detection. *Note that the framework of feature binarization driven by deep learning for efficient scene modeling is novel in the literature.*
2. We propose a blockwise binary scene model that efficiently encodes the spatio-temporal distribution information on the scene blocks. Such distribution information is measured by computing a density analysis based scoring function in the binary Hamming space, which ensures the computational efficiency of moving object detection.

2. Our method

2.1. Overview

As shown in Fig. 1, our method consists of two modules: (1) stacked denoising autoencoder (SDAE) learning; and (2) binary scene modeling based on density analysis. More specifically, we make use of deep learning to build an SDAE-based deep image representation for encoding the intrinsic structural information for a scene. For computational efficiency, we further propose a hash method to generate binary codes for effectively preserving the statistical properties of the learned feature representation in the Hamming space. Based on the hash method, we present a binary scene modeling method based on density analysis, which captures the spatio-temporal distribution information (measured by Hamming distance evaluation) for a scene in the binary Hamming space. Therefore, our method can not only learn a robust feature representation, but also construct a binary scene model based on density analysis, which results in the effectiveness and efficiency of moving object detection.

2.2. Deep feature learning by SDAE

Motivated by building an effective image representation for the robustness to complicated scene factors (e.g., noise, illumination, and local variation), we take advantage of deep learning techniques [17,18] to learn the intrinsic high-level structural scene information from a huge amount of generic natural image patches (represented by gray-scale pixels), which are generated from the Tiny image dataset [19]. Fig. 2 gives an illustration of our deep learning architecture for scene analysis.

As shown in Fig. 2(a), a 3-layer denoising autoencoder network is constructed, including the input layer, the hidden layer after encoding, and the output layer after decoding. For robustness, the original input data is typically perturbed by adding additive Gaussian noise. Subsequently, the autoencoder tries to learn a mapping function $h_{W,b}(\tilde{x}_i) \approx x_i$ to approximate and recover the original signal by establishing the low-rank hidden layer units, which encode the intrinsic high-level structural information on image patches. In mathematics, the autoencoder network is associated with the following optimization problem:

$$\min_{W,b} \frac{1}{2t} \sum_{i=1}^{t} \| x_i - \hat{x}_i \|_2^2 + \frac{\lambda}{2} \left(\| W^{(1)} \|_F^2 + \| W^{(2)} \|_F^2 \right), \tag{1}$$

where

$$\begin{aligned} \tilde{x}_i &= C(x_i), \\ y_i &= f\left(W^{(1)} \tilde{x}_i + b^{(1)}\right), \\ \hat{x}_i &= f\left(W^{(2)} y_i + b^{(2)}\right). \end{aligned} \tag{2}$$

Here, $(W^{(1)}, W^{(2)})$ and $(b^{(1)}, b^{(2)})$ denote the weight and bias variables for the encoder and decoder, respectively. $C(\cdot)$ is a function to

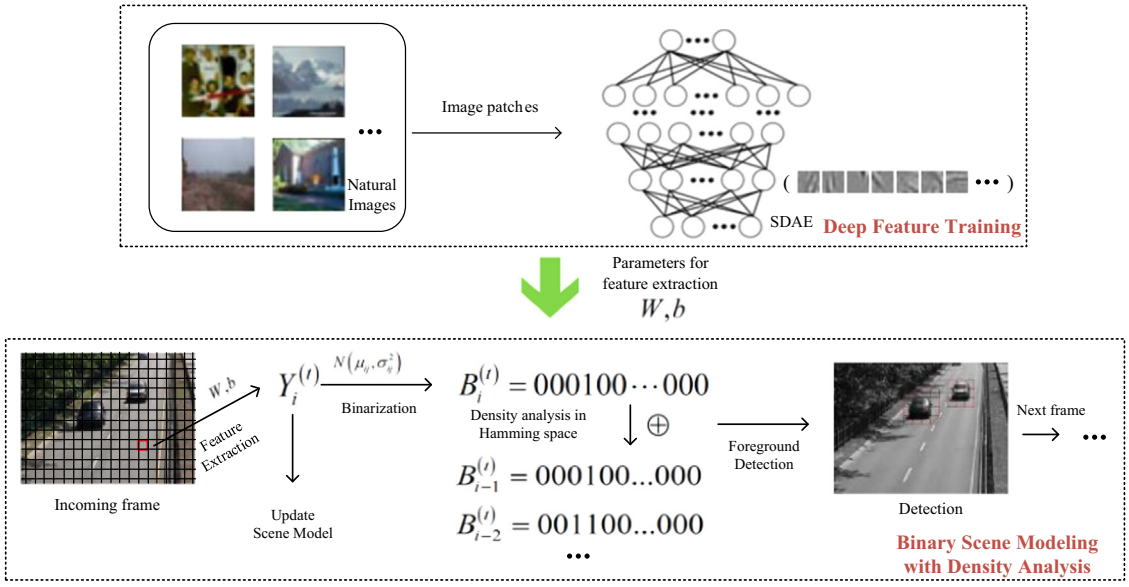

Fig. 1. The architecture of our method including the components of deep feature learning and binary scene modeling.