Pattern Recognition 67 (2017) 73–84

Contents lists available at ScienceDirect

Pattern Recognition

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

Accurate object detection using memory-based models in surveillance scenes

Xudong Li[a], Mao Ye[a,*], Yiguang Liu[b], Feng Zhang[a], Dan Liu[a], Song Tang[a]

[a] *School of Computer Science and Engineering, Center for Robotics, Key Laboratory for NeuroInformation of Ministry of Education, University of Electronic Science and Technology of China, Chengdu, 611731, China*
[b] *Vision and Image Processing Laboratory, College of Computer Science, Sichuan University, Chengdu, 610065, China*

ARTICLE INFO

Article history:
Received 17 August 2016
Revised 27 December 2016
Accepted 24 January 2017
Available online 1 February 2017

Keywords:
Convolutional neural network
Long short-term memory
Object detection

ABSTRACT

Object detection is a significant step of intelligent surveillance. The existing methods achieve the goals by technically designing or learning special features and detection models. Conversely, we propose an effective method for accurate object detection, which is inspired by the mechanism of memory and prediction in our brain. Firstly, a fix-sized window is slid on a static image to generate an image sequence. Then, a convolutional neural network extracts a feature sequence from the image sequence. Finally, a long short-term memory receives these sequential features in proper order to memorize and recognize the sequential patterns. Our contributions are 1) a memory-based classification model in which both of feature learning and sequence learning are integrated subtly, and 2) a memory-based prediction model which is specially designed to predict potential object locations in the surveillance scenes. Compared with some state-of-the-art methods, our method obtains the best performance in term of accuracy on three surveillance datasets. Our method may give some new insights on object detection researches.

1. Introduction

Intelligent surveillance is one of hot fields in computer vision due to the enormous demands for management and safety of public places. The most primary step of intelligent surveillance is object detection, especially detecting pedestrians and vehicles, which will affect the subsequent analyses, such as counting [1], tracking [2], recognition [3], re-identification [4], etc.

The traditional object detection methods technically design the hand-crafted features and the powerful object detectors, such as Haar + Adaboost [5], HOG + SVM [6], HOG + DPM [7]. Due to the fact that the hand-crafted features are usually low-level features, these traditional methods cannot obtain satisfactory results on the task of object detection. The current object detection methods are almost based on convolutional neural networks (CNNs) [8] for the sake of extracting high-level features [9]. However, most of these methods improve the detection accuracy by technically adding some layers in their models without clear guidance. Conversely, in this paper, we design our models according to some mechanisms of our brain [10].

Recently, there exists a popular consensus that our brain has a mechanism of memory and prediction [11]. When we see a vehicle at the first time, the memory mechanism is at work: our eyes scan the vehicle quickly at first; then our visual system extracts its abstract features; our brain memorizes these features at last. Thus, the memory process can be summarized as observation, abstraction and memory. Similarly, when we see a scene with several vehicles, the prediction mechanism is in action: firstly our eyes scan the whole scene rapidly; secondly our visual system extracts its abstract features; thirdly our brain gives a prediction about locations where features are similar to vehicle ones. Hence, the prediction process can be concluded as observation, abstraction and prediction.

Although the existing CNN-based methods can extract high-level features through a deep hierarchical model which shares some similarities with our visual system, there are two obvious differences between the CNN-based methods and the mechanism of memory and prediction. The first one is that the CNN-based methods take the image as the static input, but our eyes scan the image rapidly to generate an image sequence. The second one is that the CNN-based methods have no ability of sequence learning, whereas our brain contains a large amount of recurrent connections which enable to memorize the sequential patterns. With regard to recurrent neural networks, the LSTM [12] is a current great model for sequence learning due to the fact that it overcomes

* Corresponding author.
E-mail addresses: lixudong268@gmail.com (X. Li), cvlab.uestc@gmail.com (M. Ye), liuyg@scu.edu.cn (Y. Liu), xizero00@gmail.com (F. Zhang), liudan11060126@gmail.com (D. Liu), steventangsong@gmail.com (S. Tang).

http://dx.doi.org/10.1016/j.patcog.2017.01.030
0031-3203/

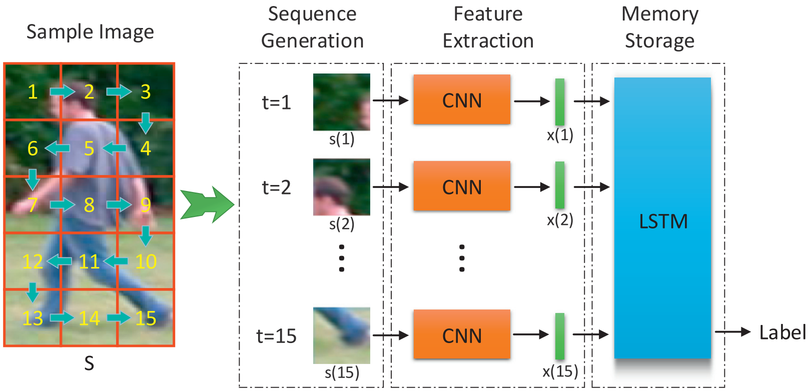

Fig. 1. The framework of our memory-based classification model.

the vanishing gradient effect in conventional recurrent neural networks. The LSTM has been proved useful for speech recognition [13], machine translation [14] and image description [15].

In this paper, according to the mechanism of memory and prediction in our brain, we propose a memory-based method for object detection in the surveillance scenes. A memory-based classification model (MCM) and a memory-based prediction model (MPM) are designed to imitate the memory process and the prediction process respectively. The MCM combines a CNN and a LSTM into an integrated framework subtly to memorize and recognize the sequential patterns (see Fig. 1), and the MPM combines a CNN and a recurrent CNN to predict the potential object locations in the surveillance scenes (see Fig. 2). Specially, the recurrent CNN is derived from the LSTM.

The main procedure of our memory-based method includes three steps. Firstly, we slide a window on the image horizontally for the sake of generating the image sequence. Secondly, we employ the CNN to extract the high-level features from the image sequence. Thirdly, the sequential features are input into the LSTM or the recurrent CNN in proper order to memorize and recognize the sequential patterns or to output the mask indicating the potential object locations.

During the training procedure, we use a simple training strategy to ease the effect of view changes in different surveillance scenes. We first train our model on a large number of samples with the horizontal view. And then we fine-tune it using a small number of samples collected from the specific surveillance scene. The experimental results show that our method obtains the best performance of pedestrian detection and vehicle detection on three surveillance datasets, when compared with some state-of-the-art methods.

The main contributions of this paper are listed as follows:

- We propose a memory-based classification model to imitate the memory process in our brain. This model aims at memorizing and recognizing the sequential patterns from the sample image by subtly combining the CNN and the LSTM into an integrated framework.
- We propose a memory-based prediction model to imitate the prediction process in our brain. This model is specially designed for object detection in the surveillance scenes. We convert the LSTM to the recurrent CNN that can output the mask indicating the potential object locations.

The remainder of this paper is organized as follows. Section 2 introduces some researches related to our method. Section 3 and Section 4 describe the details of our memory-based classification model and our memory-based prediction model respectively. Section 5 reports our experimental results, compares our method with other representative methods and makes some discussions. Section 6 concludes our paper and draws the future work.

2. Related work

In this paper, we propose the memory-based models for object detection in the specific surveillance scenes. Therefore, the most closely related techniques to our method are CNN-based object detection and scene-specific object detection.

CNN-based object detection: Many researchers pay their attention to CNNs for utilizing the ability of feature learning. Garcia et al. [16] present that the CNN with two convolutional layers can obtain higher detection rate of face detection than fully connected multi-layer perceptrons. Sermanet et al. [17] pre-train the convolutional kernels using the convolutional sparse coding. Ouyang et al. [18] learn more discriminative features through interaction with deformation and occlusion handling models. He et al. [19] propose a spatial pyramid pooling in the CNN to generate a fixed-length representation that is robust to object deformations. In order to make features more suitable for object detection, Ouyang et al. [20] pre-train the convolutional kernels with 1000-class object-level annotations instead of the image-level annotations on the ImageNet dataset [21]. Due to utilizing high-level features, these CNN-based methods have achieved better performance than the traditional methods. However, above methods encounter a problem that a large number of candidate regions need to be correctly classified.

For the sake of reducing the candidate regions, Girshick et al. propose the region-based CNN methods (R-CNN [22], Fast R-CNN [23] and Faster R-CNN [24]). R-CNN [22] employs the selective search [25] to generate the region proposals that are classified by AlexNet [26]. Fast R-CNN [23] takes an image and multiple candidate regions as input, pools these regions into the fixed-size feature maps and outputs the class labels and the bounding-box regession offsets. Faster R-CNN [24] produces a set of candidate regions by a region proposal network and makes use of Fast R-CNN for object detection. Although the region-based CNN methods raise the detection speed, their detection accuracy heavily depends on the accuracy of candidate regions, which is hard to be guaranteed in the surveillance scenes.

Another way of utilizing CNNs for object detection is to adjust the architecture of CNNs properly for the purpose of predicting accurate object locations. Szegedy et al. [27] propose a CNN-based regression model, which outputs a binary mask of object bounding boxes. Sermanet et al. [28] introduce a novel deep learning framework to predict object boundaries. Erhan et al. [29] propose a class-agnostic scalable object detection method that generates a small set of bounding boxes representing potential objects. Girshick et al. [30] formulate a deformable part model as