Chapter 9
Comparative Evaluations of Human Behavior Recognition Using Deep Learning

Jia Lu
Auckland University of Technology, New Zealand

Wei Qi Yan
Auckland University of Technology, New Zealand

ABSTRACT

With the cost decrease of security monitoring facilities such as cameras, video surveillance has been widely applied to public security and safety such as banks, transportation, shopping malls, etc. which allows police to monitor abnormal events. Through deep learning, authors can achieve high performance of human behavior detection and recognition by using model training and tests. This chapter uses public datasets Weizmann dataset and KTH dataset to train deep learning models. Four deep learning models were investigated for human behavior recognition. Results show that YOLOv3 model is the best one and achieved 96.29% of mAP based on Weizmann dataset and 84.58% of mAP on KTH dataset. The chapter conducts human behavior recognition using deep learning and evaluates the outcomes of different approaches with the support of the datasets.

INTRODUCTION

With the rapidly increasing number of surveillance cameras in various scenes, locating the region of interest (ROI) from thousands of surveillance videos has become one of the prominent problems. On the other hand, with continuous expansion of surveillance systems, a vast amount of video footages has been archived, which become more and more tough to find useful information from these large data, identifying a target is also less efficient. Therefore, how to make surveillance more efficiency in the environment with tremendous big visual data needs to be taken into consideration.

DOI: 10.4018/978-1-7998-2701-6.ch009

For traditional human behavior recognition, both low-level (without semantic understanding) and high-level processes (with semantic understanding) are needed to be considered. The low-level process helps to locate the region of interest and reduce the useless information from video footages; meanwhile, the high-level process will analyze these features to achieve the behavior recognition ultimately.

Nowadays, with the increasing capacity of computing devices, Deep Neural Networks (DNNs) obtained a massive attention to detect objects, which lead to a new era of computer vision (Lu et al., 2017). Deep learning models (Hinton et al., 2006) contain multiple hidden layers, pre-training methods are adopted to alleviate the troubles of local optimal solution, the hidden layers are up to many with the "depth", thus it is called "Deep Learning". The state-of-the-art methods mainly rely on artificial neural networks, such as Convolutional Neural Networks (CNNs or ConvNets), R-CNNs, Fast R-CNNs, Faster R-CNNs, Mask R-CNN, and SSD (Single Shot Multi-Box Detector). Moreover, deep learning has been implemented in both supervised or unsupervised models (Ji et al., 2013). Apparently, the work has clearly shown the difference between deep neural networks and shallow neural networks as well as conventional machine learning in various aspects (Liu et al., 2016).

In human behavior recognition, bounding box was proposed for deep neural networks to resample these proposed pixels. Because there are inherent local patterns in an image such as eyes, nose, mouth, etc., CNN is derived by combining digital image processing and artificial neural networks, which links the upper and lower layers through the convolution kernels. The convolution kernels are shared among all images, the images still retain the original position after the convolution operations. Compared to traditional approaches, deep learning conducted in an end-to-end process along with more higher classification probability which also shows an invariant to illumination, pose, etc. (LeCun et al., 2004). Moreover, region proposal methods and region-based convolutional neural networks (R-CNN) are more successful with high precision in pattern recognition (Ren et al., 2017). The selective Faster R-CNN (Uijlings et al., 2013) also carries out outperformance in a series of comparisons (Ren et al., 2017; He et al., 2016). Region proposal networks (RPN) were trained for the end-to-end purpose and generated the accurate location of the regions of interest. RPN and Fast R-CNN have been merged together into a single network which is able to accelerate the detection (Ren et al., 2017).

This book chapter is organized as follows. Literature review is presented at Section 2, our method is described in Section 3. Our results are demonstrated in Section 4, the conclusion is drawn in Section 5.

RELATED WORK

Human behavior understanding refers to analyze and recognize human motion patterns, and describe it with natural languages (Aggarwal et al., 1997). The motion sequence can be considered as the traversal process of static actions in different state nodes (Guo et al., 1994). The joint probability of traversal process is therefore calculated, its maximum value is taken into consideration for classification (Fujiyoshi et al., 2004).

Human behavior recognition is to identify the predefined behaviors automatically so as to reduce human labor effectively (Aggarwal et al., 1997). To recognize and analyze the periodic motions of human behavior, spatial-temporal model (Rui et al., 2000) and periodic model (Cutler et al., 2000) were proposed. Human behavior recognition can mainly split into twofold which contains template-matching-based methods and space-state-based methods.