Engineering Applications of Artificial Intelligence 67 (2018) 226–234

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Deep convolutional framework for abnormal behavior detection in a smart surveillance system

Kwang-Eun Ko, Kwee-Bo Sim *

School of Electrical and Electronics Engineering, Chung-Ang University, 84, Heukseok-ro, Dongjak-gu, Seoul, 06974, Republic of Korea

ARTICLE INFO

MSC:
00-01
99-00

Keywords:
Behavior recognition
Convolutional neural network
Long short-term memory
Smart surveillance system

ABSTRACT

The ability to instantly detect risky behavior in video surveillance systems is a critical issue in a smart surveillance system. In this paper, a unified framework based on a deep convolutional framework is proposed to detect abnormal human behavior from a standard RGB image. The objective of the unified structure is to improve detection speed while maintaining recognition accuracy. The deep convolutional framework consists of (1) a human subject detection and discrimination module that is proposed to solve the problem of separating object entities, in contrast to previous object detection algorithms, (2) a posture classification module to extract spatial features of abnormal behavior, and (3) an abnormal behavior detection module based on long short-term memory (LSTM). Experiments on a benchmark dataset evaluate the potential of the proposed method in the context of smart surveillance. The results indicate that the proposed method provides satisfactory performance in detecting abnormal behavior in a real-world scenario.

1. Introduction

Background. The Fourth Industrial Revolution, highlighted at the World Economic Forum in Davos, Switzerland, on 20 January 2016, predicted that the combination of artificial intelligence (AI) technology and diverse industrial fields will create fundamental changes in multidisciplinary areas. These changes are based on the exponentially advanced recent AI-related technologies, such as computer vision, robotics, and machine learning. Examples of the innovative next-generation technologies based on this trend include ubiquitous, mobile supercomputing, intelligent robotics, and self-driving cars, which can dramatically change the way people live (Schwab, 2017).

This research topic is closely related to state-of-the-art HCI (human–computer interaction) applications. In the most common user interface scenario, when a computer interacts with a human, it performs an operation corresponding to the command, which is transferred through a standard interface, such as a keyboard, mouse, and touchscreen. Recent HCI works have attempted to develop an approach that enables more intuitive interactions between a human and a computer through a computer vision framework. They tried to exclude as much as possible that use separate interface devices for interaction, such as the human visual recognition process. It is widely believed that one of the next shifts in HCI will be the endowing of computers with the capability of understanding human behavior through computer vision technology. This topic is increasing in popularity across the academic and industrial fields involving security/surveillance (Vishwakarma and Agrawal, 2013), industrial robotics (Roitberg et al., 2015), and affective computing (Kleinsmith and Bianchi-Berthouze, 2013). This trend has encouraged the computer vision community to address issues related to overcoming technical limitations from existing approaches based on the standard RGB camera.

At the current level of computer vision technology, it is possible to acquire various types of image data through diverse vision sensors, such as a stationary camera and a stereo camera, and therefore it is becoming more convenient for the acquired scene to be aware of context such as the spatiotemporal state transitions of objects (e.g., human behavior). In particular, the recent emergence of price-competitive RGB-D sensors has made a breakthrough in the deep-seated issues of computer vision, such as the subtraction of background and the elimination of disturbance from the light source or other occlusions. The combination of these types of depth sensors and powerful pattern recognition algorithms aids in increasing the accuracy and speed of understanding human

* Corresponding author.
E-mail addresses: kkeun@cau.ac.kr (K.-E. Ko), kbsim@cau.ac.kr (K.-B. Sim).

https://doi.org/10.1016/j.engappai.2017.10.001
Received 18 June 2017; Received in revised form 15 September 2017; Accepted 9 October 2017
0952-1976/

behavior (Vieira et al., 2012; Li and Leung, 2017; Slama et al., 2015; Chen et al., 2016). However, these approaches have only been implemented in the laboratory and are not as precise as the human motion analysis based on 3D data from expensive marker-based motion capture systems (Barbič et al., 2004; Moeslund et al., 2006). They are also difficult to apply in a practical industrial field because of limitations in the characteristics inherent to depth sensors such as effective distance measurement limitations and noise sensitivity caused by environmental changes. To develop a practical application, such as for surveillance system s, it is essential to advance the behavior recognition technology using RGB images of standard quality.

Previous studies on human behavior recognition started at the level of recognizing postures/gestures expressed in still images. As a result, it was possible to recognize simple motions such as walking, running, and sitting. More recent studies progressed to the point of estimating the purpose of the behavior based on the motion information (Poppe, 2010). However, actual human behavior is closely connected with the surrounding environment and other objects in addition to the motion information (Dedeoğlu et al., 2006). For example, if an individual performs two separate behaviors (e.g., making a phone call and drinking coffee) having similar motor patterns, the behavioral class of each motion will depend on the information about the object that is the target of the motion. In addition, understanding group actions performed by people with close interactions (e.g., push, hug, and high five) has become one of the major challenges in recent human behavior recognition research (Kong and Fu, 2016; Huynh-The et al., 2016). Therefore, in order to recognize behavior more precisely, state-of-the-art studies insist that it is necessary to develop the ability of contextual awareness of the relationship between environment and objects (Abowd et al., 1999).

This paper presents a technology for detecting abnormal behaviors that can occur during close interactions between people. The proposed approach is applicable for intelligent monitoring/surveillance of the elderly population in preparation for an aging society. In order to consider the safety of elderly people suffering from degenerative diseases, such as dementia, Alzheimer's disease, and Parkinson's disease, a round-the-clock surveillance system capable of identifying when an individual performs an abnormal behavior is required.

Organization. The organization of this paper is as follows. Section 2 addresses previous studies related to the proposed abnormal behavior detection method in the surveillance process and the capabilities of current technologies. Section 3 describes core technologies that constitute the proposed method and the detailed procedures of the proposed method. Section 4 describes the experiments performed to verify the performance of the proposed method and evaluate the results. Finally, the conclusion and suggestions for future works are presented.

2. Related works

Surveillance event detection. One of the important challenges of visual surveillance is the autonomous and online detection of events caused by the performance of abnormal behaviors by objects under surveillance. There are a number of steps for event detection in the autonomous surveillance process. First, real-time object detection is required. For example, some literature asserts that detecting regions in the video that correspond to objects, such as people, should be handled by comparing spatial appearance features. The approach presented in Dedeoğlu et al. (2006) extracts silhouettes of the moving objects using an adaptive background-subtraction scheme to detect the spatial region of the objects; then, a template-matching algorithm based on the silhouette feature is executed to classify objects corresponding to humans. The next step is to find the proper temporal region corresponding to the target event from within the entire video that is acquired from the surveillance process. The method proposed in Şaykol et al. (2010a) focuses on keyframe labeling for event classification based on the result of the behavior appearance recognition. In this approach, keyframes are labeled by detecting the temporal region where the behavior occurs within the input frame sequence. The event classification is then accomplished by representing the input stream as a temporally ordered sequence of keyframe labels. In addition, an autonomous video surveillance system should support both online semantic analysis and offline inspection. There are a number of studies on query processes to retrieve events and objects from surveillance video. For example, the system described in Şaykol et al. (2010b) performs scenario-based query processing for an archive of surveillance system. This query processing system has become a useful mechanism for effective offline inspection, such as after-the-fact activity analysis.

Deep learning. Recently, machine learning studies have developed various algorithms based on deep learning that exhibit remarkable capabilities in computer vision tasks, as for example convolutional neural networks (CNNs), which achieved the best accuracy in the object recognition and detection task with large-scale image databases in the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) 2012 (Everingham et al., 2015) and the PASCAL VOC (Visual Object Classes) challenge (Everingham et al., 2015). The original CNN architecture proposed by Krizhevsky et al. is called AlexNet (Krizhevsky et al., 2012), and its top-5 error rate was 15.3%, which exceeded the result of the second-place team by more than 40%. The algorithm outperformed previous approaches such as support vector machine (SVM) and traditional pattern classification algorithms. Furthermore, the development of support hardware, such as a graphics processing unit (GPU), along with the emergence of the concept of big data, makes it possible to learn and evaluate a large-scale deep learning architecture in a short time, and various attempts have focused on this. High-performance GPU processing (such as with CUDA) and large public image repositories (such as ImageNet (Deng et al., 2009)) have enabled CNNs to become the most popular method in the computer vision field. Several attempts have focused on enhancing the performance of AlexNet to achieve improved accuracy in image- and video-based object recognition tasks. The best-performing submissions to the ILSVRC were ZF-Net (Zeiler and Fergus, 2014) in 2013, GoogLeNet (Szegedy et al., 2015) in 2014, and ResNet (He et al., 2016) in 2015. Additionally, VGGNet has been one of the best-performing models in the image recognition challenge (Simonyan and Zisserman, 2014b). The architecture of VGGNet is relatively simple compared with those of other winning models and is easy to deform. Given this property, VGGNet is convenient to use even for video recognition and exhibits an excellent recognition rate.

Behavior recognition. Automatic recognition of behavior performed by a human object is essential for building an intelligent surveillance system. Behavior recognition for the purpose of surveillance event detection refers to the classification of a specific behavior pattern from a continuous input stream that consists of an individual's motion elements and context elements. Motion information in particular is widely utilized for a variety of practical applications (Ni et al., 2011). Most computer vision studies have conducted the behavior recognition at the level of motion analysis using standard RGB images or depth images. These previous computer vision approaches have followed a few formalized steps for human behavior recognition (Schmidt, 2000). Extracting the optimal motion features from image frame sequences representing a behavior is the first challenge. Most of the existing approaches have tried to extract motion features of the observed behavior by using optical flow, which is extracted from adjacent image frames as key feature data representing a distinct pattern (Chaudhry et al., 2009). Recently, CNN-based approaches integrating feature extraction and classification of the motion of behavior have attracted much attention