Vision-based Toddler Physical Activity Recognition using Deep Learning

Norasyikin Fadilah[1], Mohd Zamri Ibrahim[2*], Rosdiyana Samad[3]*

[1,2,3]*Faculty of Electrical & Electronics Engineering Technology, Universiti Malaysia Pahang, Pahang, Malaysia*

[1*]norasyikin@ump.edu.my, [2*]zamri@ump.edu.my

Keywords: activity recognition, deep learning, LSTM, 2D skeleton

Abstract

Human activity recognition (HAR) is a system for understanding human movements and behaviour. It has been applied in many fields such as video surveillance, behaviour analysis, and human-computer interaction. The state-of-the-art studies on HAR generally focus their attention on public dataset which mostly consist of adults as their subjects. Research on HAR for children especially toddlers is important to facilitate their surveillance by monitoring their activities automatically. Since toddlers possess different anatomical proportions than adults, their unusual movements can be a challenge to infer. In this paper, a vision-based deep learning HAR system for toddlers was developed based on skeleton features. Videos of toddlers' activities in a day-care were obtained through different public sources. 2D skeleton data were then extracted from every frame of these videos using a pre-trained deep learning network. These skeleton data were trained on LSTM and fully connected network to infer the toddler's activities. Results showed that this proposed framework managed to achieve 75% accuracies for three toddlers' activities which are jumping, sitting, and standing.

1 Introduction

Visual observations and analysis of children's natural behaviours are useful to the early detection of developmental disorders, health indicators and facilitation of their surveillance. Many places such as private homes, day care centres or healthcare centres are equipped with cameras for surveillance purpose. This makes the videos for children behaviours observations more easily accessible and the implementation of automatic children's actions possible. There are ongoing research that study on developing children action recognition for specific purposes based on video data. Examples of the purposes include detection of typical behaviours of children with autism spectrum disorder (ASD)[1] or cerebral palsy[2], recognizing health levels in children[3], measurement of physical activity[4], and determining gross motor skills[5]. Most of these applications involve the processing of videos of performing different types of activities, which generally refer to human activity recognition (HAR).

HAR identifies human activities or behaviours through a series of observations between the actions and environmental conditions. It is a system that is applicable in variety of domains including surveillance, healthcare, human-computer interaction, behaviour analysis, and remote activity monitoring. Traditionally, the task of HAR has been observed manually by humans through the face-to-face observations or analysing videos by their interpretations. The traditional method is subjective, thus may lead to errors. Moreover, manual analysing of videos will delay prompt action taken especially when it involves the safety and development of children. For example, in current surveillance system for a day care centre, any misconduct of the care person, accidents or abnormal behaviour of the children will only be realized through reports and manual checking of the recorded videos. Therefore, automated HAR for this purpose is important and has becoming a dynamic active research topic nowadays. Due to exceptional development of sensor technology, researchers have been using wearable sensors, ambient sensors, or cameras to be integrated in HAR systems[6]–[8]. From the outputs of these sensors, important features are acquired, processed, and then extracted by computing numeric or symbolic information to establish a system that can recognize the human activities.

In computer vision field, the goal of HAR is to analyse a video to identify the actions taking place by one or multiple persons. A video is processed frame by frame to extract the spatial features to represent every human activity, followed by developing the activity recognition algorithm [9]. The feature vector to represent human action can be in the form of raw RGB image, optical flow, trajectory, depth, or pose. HAR models using raw RGB data is the most complicated process because they need to extract and interpret complex features. Most