Contents lists available at ScienceDirect

CIRP Annals - Manufacturing Technology

journal homepage: http://ees.elsevier.com/cirp/default.asp

Deep learning-based human motion recognition for predictive context-aware human-robot collaboration

Peng Wang [a], Hongyi Liu [b], Lihui Wang (1)[b], Robert X. Gao (1)[a,*]

[a] Department of Mechanical and Aerospace Engineering, Case Western Reserve University, Cleveland, OH, USA
[b] Department of Production Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

ARTICLE INFO

Keywords:
Motion
Predictive model
Machine learning

ABSTRACT

Timely context awareness is key to improving operation efficiency and safety in human-robot collaboration (HRC) for intelligent manufacturing. Visual observation of human workers' motion provides informative clues about the specific tasks to be performed, thus can be explored for establishing accurate and reliable context awareness. Towards this goal, this paper investigates deep learning as a data driven technique for continuous human motion analysis and future HRC needs prediction, leading to improved robot planning and control in accomplishing a shared task. A case study in engine assembly is carried out to validate the feasibility of the proposed method.

1. Introduction

In recent years, human-robot collaboration (HRC) has emerged as a key technology for intelligent manufacturing. Instead of strictly separating human operators and robots due to safety reasons, HRC allows the humans and the robots to work together in the same work space and collaborate in performing the same tasks [1,2]. In an HRC system, the robots are expected to actively assist human operators in performing complex tasks, besides independently performing their own tasks. While ensuring safety is a major objective, an HRC system aims at improving the operation efficiency and productivity. For this purpose, the robots are required to track a human operator's motion, identify the context of collaboration [3], and predict how the human operator would behave subsequently to accomplish a certain task. Identification of the context of collaboration involves identification of the objects (parts or tools) that the human worker is handling, the sequences of performing actions during the tasks, and the work space environment. With context awareness, the robots will be able to know how it can effectively assist the human operator, e.g. when to pass what tools or parts to the human operator to improve productivity of the HRC system while maintaining safety [4]. As significant variability and heterogeneities may exist among human operators when performing the same task, the context of collaboration may vary accordingly. Therefore, context awareness is as important as human motion recognition for establishing reliable HRC systems. In this paper, the combination of human motion recognition and context awareness is regarded and termed as human action recognition.

Motion recognition is a crucial part for establishing an effective HRC system [5]. With the human operator's motion accurately tracked, real-time human motion in the HRC environment can be accomplished. There have been reported studies on recognizing and understanding the motions of humans [6–8]. In these studies, traditional machine learning methods such as random forest [6], Gaussian mixture models (GMM) [7], and neural networks [8] were applied to recognize and understand human motion. Many of the traditional machine learning methods have reported to achieve a 70–80% accuracy on motion recognition. Based on the understanding of human operator's intention revealed by the recognized motion, some studies tried to predict the human operator's future motions for robot planning in an HRC environment [9–11]. Hidden Markov model (HMM) [9], Bayes network [10], and Bag-of-words [11] etc., have been used to set up statistical models for human motion prediction. Due to the relatively low accuracy in motion recognition, the accuracy of motion prediction rarely exceeds 80%. In addition, these studies did not take into consideration of the collaboration context, and they cannot ensure reliable estimation of human intention for robot planning.

Deep learning has emerged as a new machine learning architecture with significant capability in discovering and learning complex patterns underlying a large amount of data. As such, it provides a new approach to improving the recognition accuracy of human motions. Recent reports have shown that deep learning can outperform human experts in recognition or strategy-related tasks [12,13]. Compared with traditional machine learning techniques, the structure of deep learning networks involves multiple hidden layers to enable the extraction of features that are deeply embedded in the data, forming abstract concepts in a hierarchy manner [14]. So far, deep learning has been successfully demonstrated in several application domains, including image recognition, speech recognition, and data analysis [14].

This paper presents research on deep convolutional neural network (DCNN) to recognize human motions and identify the context of associated action for accurate and robust inference of human operator's intention in performing manufacturing tasks. A well-defined DCNN structure, the AlexNet, has been modified through

[*] Corresponding author.
E-mail address: Robert.Gao@case.edu (X. Gao).

https://doi.org/10.1016/j.cirp.2018.04.066
0007-8506/© 2018 Published by Elsevier Ltd on behalf of CIRP.

a transfer learning-enabled tuning method, to improve the rate of recognition of human operator's actions. The established deep learning context-aware human motion recognition model is experimentally evaluated for an automotive engine assembly process.

2. Human action recognition in HRC

In an HRC manufacturing system, human operators and robots team up and collaborate on completing complex tasks, in a diverse range of scenarios with highly dynamic and uncertain shop floor environments. The robots are expected to assist human besides independently performing tasks. The goal of the collaboration is to: (1) ensure safety in the collaborative work space and (2) increase production efficiency. For this purpose, the robots should be able to accurately capture the human operator's actions and understand their intentions, while taking into consideration the variability and heterogeneities among human operators in performing the same tasks. For example, to collaboratively completing an assembly task, human actions (e.g. placing a part at a certain location or driving screws) and the context of the actions are captured by video cameras, and the associated videos/images are analyzed to extract the information needed for robot task planning, such as when to pass what tools to human operator, as illustrated in Fig. 1.

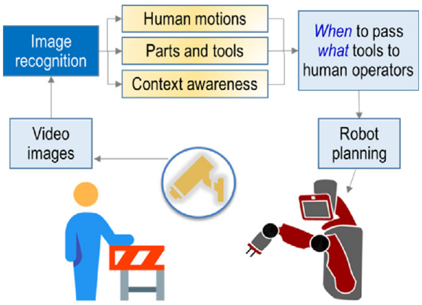

Fig. 1. Role of human action recognition in HRC.

Human body motions associated with certain tasks may be similar regardless of the context of the tasks. For example, there could be no distinct difference between body motions when grasping a part and a tool (e.g., a screwdriver). In an HRC system, human actions are first recognized in terms of generic body motions (e.g. standing, grasping, and holding). After motion recognition is performed, the context of the actions are recognized to assist in the identification of the operator's intention (see Fig. 2). This is aimed at helping the robot understand what specific actions the human operator intends to perform so that the robot can assist correspondingly. For example, when the robot captures the scenario of a human holding a screwdriver, the robot would recognize that the human intends to drive screws. As a response, it would fetch a screw and pass it on to the human operator.

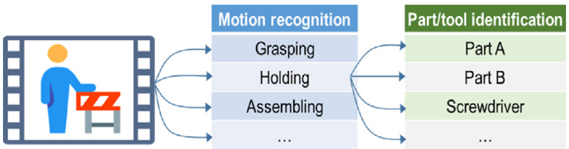

Fig. 2. Steps for human action recognition.

In this paper, deep learning is investigated to process video images for human action recognition. Each video is processed frame by frame to develop a sequential order of actions that a human operator would take to complete a specific task. The process of analyzing each frame consists of two steps: human motion recognition and part/tool identification. These are implemented by two separate deep neural networks. To train the networks, images of human motions and parts/tools associated with the tasks are captured first. Since the background where the video images were taken may be "noisy", with multiple objects involved, reliably identifying parts/tools that the human operator was working with can be challenging. A solution to this problem is to take a two-step approach, by (1) identifying the human motions associated with the task and categorize them into representative categories (such as grasping, holding, assembling), and (2) specifying from images in the "holding" category what specific parts/tools the human operator was holding. This is the approach taken in this study.

3. DCNN and AlexNet for image recognition

3.1. Architecture of DCNN

Deep convolutional neural network (DCNN), as one of the deep neural network structures, is developed specifically for image processing and recognition. It consists of hierarchically arranged trainable layers that automatically detect and learn patterns underlying the images. Along the progression of layers, multi-level representation of image features are extracted, from low-order features (e.g. edge and color) to high-order features (e.g. domain-specific motif and object).

A typical DCNN is comprised of one or more *convolutional*, *pooling*, and *fully connected* layers. Given a 2D image, features are extracted through kernel-based convolution operation in the convolutional layers:

$$x_j^l = \phi\left(b_j^l + \sum_{i=1}^{M} x_i^{l-1} * k_{ij}^l\right) \tag{1}$$

where a pixel (or a neuron) j in the l^{th} convolutional layer is obtained as the weighted sum of M selected pixels from the preceding l-1th layer, after a bias b and a nonlinear activation ϕ are added. A kernel is specified by the kernel size M and kernel functionality that is defined by the particular weights $k_{1:M}$. The kernel weights are either trained or pre-defined to detect particular features (e.g. curve and edge). For each *convolutional layer*, multiple kernels can be used to extract multiple features of interest, with the convolutional results termed as feature maps. Effective feature selection at successive layers that distinguish between different image categories is critical to achieve accurate image recognition and classification.

A *pooling layer* is usually appended to a *convolutional layer*, which fuses certain number of pixels in the feature maps into one pixel by calculating the average or maximum values among the pixels, to reduce the amount of neurons and weights involved in the network and improve the computational efficiency. As another benefit, pooling alleviates the overfitting problem. By executing pooling, relative locations of features instead of exact locations are analyzed, for recognition of images with translation and rotation properties.

Taking the highest order features from the last *convolutional layer* as the inputs, the *fully connected layers* perform image classification. The architecture of DCNN (e.g. number of layers and number of feature maps in one layer) determines the depth of image decomposition to discover the underlying patterns, which significantly affects the image classification accuracy. In this paper, a well-defined DCNN architecture, *AlexNet*, is adopted as the basic deep learning network structure. It was modified through a transfer learning-enabled algorithm to improve the learning ability for human action identification.

3.2. AlexNet

AlexNet is developed by Alex Krizhevsky et al. to classify 1.2 million high resolution images into 1000 different classes in the ImageNet LSVRC contest [15]. AlexNet demonstrated superior performance in image recognition by reducing the classification error rate from 25.8% to 16.4%, which represents a 36% improvement.

As shown in Fig. 3, AlexNet consists of five convolutional layers, three pooling layers, three fully connected layers, and 650,000 neurons in total. Over 60 million parameters needed to be trained for image recognition. For example, the first convolutional layer C1 contains 96 feature maps of 55 × 55 in dimension, or 55 × 55 × 96 = 290,400 neurons in total. Each neuron is generated from a 11 × 11 × 3 kernel