Chapter 32
Multi-Sensor Motion Fusion Using Deep Neural Network Learning

Xinyao Sun
University of Alberta, Canada

Anup Basu
University of Alberta, Canada

Irene Cheng
University of Alberta, Canada

ABSTRACT

Hand pose estimation for a continuous sequence has been an important topic not only in computer vision but also human-computer-interaction. Exploring the feasibility to use hand gestures to replace input devices, e.g., mouse, keyboard, joy-stick and touch screen, has attracted increasing attention from academic and industrial researchers. The fast advancement of hand pose estimation techniques is complemented by the rapid development of smart sensors technology such as Kinect and Leap. We introduce a hand pose estimation multi-sensor system. Two tracking models are proposed based on Deep (Recurrent) Neural Network (DRNN) architecture. Data captured from different sensors are analyzed and fused to produce an optimal hand pose sequence. Experimental results show that our models outperform previous methods with better accuracy, meeting real-time application requirement. Performance comparisons between DNN and DRNN, spatial and spatial-temporal features, and single- and dual- sensors, are also presented.

INTRODUCTION

Hand gesture is a powerful communication mechanism and hand pose estimation has been a popular research topic in the area of computer vision, image processing and Human Computer Interaction (HCI). While sign language plays an essential role in human-to-human communication, hand gestures have often been used in HCI applications, replacing a variety of traditional input devices (e.g., mouse, keyboard,

DOI: 10.4018/978-1-7998-0414-7.ch032

joystick, touch pad, and touch screen). The use of hand movements allows users to interact with computer interfaces directly and naturally without the restriction imposed by mechanical devices. There are more and more smart-sensor based contact-free input systems available to enhance this touch-free experience. Compact smart sensors have become consumer products and are commonplace for home entertainment systems (Blaha & Gupta, 2014). High precision sensors also benefit medical and clinical applications, where touching mechanical input device requires a re-sterilization step before continuing the operation (Ebert, Flach, Thali, & Ross, 2014) (Chng, 2012). Touch-free input systems based on hand gestures have advanced HCI technology to a new era and their performances are improving.

Motion detection based on smart sensing is commonly seen in consumer applications (e.g., games). The trend of touch-free HCI is strongly supported by two technologies. First is the release of new optical sensors such as Microsoft's Kinect and Leap Motion Controller (LMC). Kinect is able to provide both color and depth information by using an RGB camera and a depth sensor. The Kinect technology has been successfully applied to analyze rigid human body movement. Researchers (Tang, Lu, Wang, Huang, & Li, 2015) (Oikonomidis, Kyriazis, & Argyros, 2011) combine both color and depth information to precisely detect the location of human hands and predict joint positions, and then construct a hand skeleton following an Inverse Kinematics process (Sridhar, Mueller, Oulasvirta, & Theobalt, 2015). Compared to Kinect, LMC is a higher precision but shorter detection range consumer-grade Infrared Radiation (IR) based optical sensor developed by Leap Motion Inc. LMC is primarily designed for hand gesture and hand finger position detection. LMC offers precision at millimeter level and is able to capture fully articulated hand skeleton at a sampling rate of over 120 per second. Both Kinect and LMC sensors enhance the 2D hand motion image features with a third dimension – depth information. The depth data accelerates hand detection and increases the accuracy of hand pose estimation.

The second complementing technology is the rapid development of deep learning algorithms. Deep Neural Network (DNN) has shown its capability in feature extraction by simulating the functionalities of the human brain with layers of neurons. A classical DNN model can be used to automatically extract and analyze features from raw data. The model has been successfully applied in many pattern recognition problems relating to speech signals and images (Lic, Seide, & Yuc, 2011) (Seide, Li, & Yu, 2011). Convolutional Neural Network (CNN), as a standard architecture of Deep Learning, shows excellent performance when solving many computer vision problems. Researchers applied CNN to predict the 3D joint locations of a hand with a depth map, or with a depth map and 2D RGB image combined (Oberweger, Wohlhart, & Lepetit, 2015) (Tang, Jin Chang, Tejani, & Kim, 2014).

Despite these technological advances of hardware and algorithms, accurately tracking the articulation of hand poses in real-time is still an open research problem. A sensor has a limited viewing volume and there is noisy sensor data caused by the complex and fast finger movements (Von Hardenberg & Bérard, 2001). The uncontrolled environment can lead to possible background cluttering, which adversely affects image-based detection techniques (Chen, Georganas, & Petriu, 2007). Too many degrees of freedom of the human hand push the pose analysis process into a very high dimensional space. Computational complexity for a fully articulated hand model is often high, leading to real-time performance issues. Fingers' similarity and out-of-plane transformation lead to full or partial occlusion with each other (Rossol, Cheng, & Basu, 2016). The limited and static viewpoint of a standalone sensor reduces the effective user interaction space (Supancic, Rogez, Yang, Shotton, & Ramanan, 2015) in HCI applications.

In this work, we focus on improving hand pose estimation and enhancing existing single-view based tracking systems. We propose using multiple sensors and fuse the estimations from multiple sources in order to predict hand pose motion more accurately. We apply Deep Learning techniques to intelligently

Related Content