A Hybrid Feature Extraction Approach for Human Detection, Tracking and Activity Recognition Using Depth Sensors

Shaharyar Kamal1 · Ahmad Jalal2

Received: 21 October 2014 / Accepted: 15 November 2015 / Published online: 30 November 2015 © King Fahd University of Petroleum & Minerals 2015

Abstract This paper presents spatiotemporal hybrid features, human tracking, and activity recognition into a single framework from video sequences captured by a RGB-D sensor. Initially, we received a sequence of depth maps to extract human silhouettes from the noisy background and track them using temporal human motion information from each frame. Then, hybrid features as optical flow motion features and distance parameters features are extracted from the depth silhouette region and used in an augmented form to work as spatiotemporal features. In order to represent each activity in a better way, the augmented features are being clustered and symbolized by self-organization maps. Finally, these features are then processed by hidden Markov models to train and recognize human activities based on transition and emission probabilities values. The experimental results show the superiority of the proposed method over the state-of-the-art methods using two challenging depth images datasets.

Keywords Depth images · Hybrid features · Activity tracking · Recognition system

1 Introduction

Human detection, tracking and activity recognition are the three most vital fields in computer vision and widely applied in a number of real-world applications, e.g., human–computer interaction, analyzing of human motion, 3D games, detecting uncertain events, video surveillance and activity recognition [1–5]. In the past decades, research on these fields mostly used optical body markers, conventional digital cameras and other motion sensors [6–8]. With the advancement of imaging devices [9,10], the recent development of low-cost and easy operating video sensors (i.e., Kinect) provide major contributions in visual computing tasks. In addition, depth maps produce efficient representation of human shape silhouettes, recover skeleton joints from a single depth frame, insensitive to light change and distance information based on pixel intensities [11]. However, video-based activity recognition systems start with removing background noise from each frame, extracting human silhouettes and deal with feature representation techniques [12,13] to produce significant accuracy and computational efficiency during training/testing each activity.

Depending on the feature extraction and representation, the conventional algorithms are classified into two types: template-based approach and model-based approach. In template-based approach, each activity is represented as human silhouettes or activity shape information such as pixel points or optical flow. In [14], hierarchical motion history mechanism is proposed to compute both global and dense (local) motion flow from the motion history images (MHI). MHI pyramid enables efficient calculation of image motions using fixed-size gradient operators, whereas to analyze motion field in particular activity, polar histogram of motion orientation is used which makes it computationally inexpensive method for recognizing human behavior.
In [15], translation and scaling invariant feature system is designed where 2D feature maps are computed first through Radon transform of full-human silhouettes followed by computing 1D profile through R transform. These features are applied for dimension reduction and passed through discrimination process to make the features more prominent. However, depth features act as spatiotemporal properties which are trained/tested via recognizer engine. On the other hand, in case of model-based approach, human body is replaced by specific skeleton-based definition having joint points information used for features purpose. In [16], a novel feature approach is proposed based on position differences of joints and eigenjoints which provides human behaviors information including static postures, motion and offset. These features deal with Naive-Bayes-nearest neighbor classifier for multi-class action recognition followed by support vector machine for recognition. In [17], real-time life logging system is designed which provides software routine to recognize human behavior and activities by extracting body joints information from depth silhouettes. These joints are used for feature generation to get motion parameters and trained/recognized activities using hidden Markov model (HMM). Relying on these both template and model-based methods [14–17] proved more efficient recognition accuracy using human silhouette. However, it is still difficult in complex activities where body parts are not cleared or hidden. Also, template model faces limited number of joint points information during self-occlusions which directly affect recognition accuracy. Therefore, we needed full-body silhouettes and skeleton-based information to be merged for efficient feature extraction and activity recognition.

In this work, we focus on recognizing human activities by considering a sequences of depth maps utilizing the Microsoft Kinect camera. To extract human depth silhouette from noisy depth data, we detect and track human silhouettes based on background subtraction and floor removal process. Then, hybrid feature extraction technique is applied which includes template-based optical flow and distance parameters features. Finally, self-organized map (SOM) is applied on the augmented features for clustering and combining with HMM for training/testing human activities. In the extensive experiments, we achieved recognition accuracy superior to all state-of-the-art methods on two depth datasets, i.e., our annotated depth dataset and MSRAction3D. One of the aims of the proposed system is to use it in a wide range of applications such as healthcare system to monitor and recognize activities of patients at daily basis, video surveillance to observe uncertain events and humanoid robotic systems allowing robots to act as human beings.

The rest of the paper is summarized as follows. Section 2 outlines the proposed architecture of our system where depth maps pre-processing, feature extraction based on full-body silhouettes along with segmented body joints information and activity training/testing using HMM are discussed. Section 3 focuses on experimental results as well as recognition accuracy by considering proposed and state-of-the-art methods. Section 4 summarizes the conclusion.

2 Proposed System Methodology

In this work, we start with the pre-processing of the incoming raw depth data captured by depth video sensor. Then, human silhouettes extracted from the time-sequential activity video images are processed by hybrid feature technique having optical flow information and distance parameter features. Finally, quantify the depth silhouettes using SOM with respect to best matching unit and find maximum likelihood after training each activity using recognizer engine. Figure 1 shows the general framework of proposed human activity recognition system.

2.1 Depth Maps Acquisition and Pre-processing

The major factor of depth map acquisition and pre-processing is to find the depth human silhouettes from the noisy background using depth camera (i.e., PrimeSence and bumblebee). These cameras deal with both RGB images and depth raw data, while the depth data contain the intensities of pixels having gray level values [18]. Figure 2 illustrates some examples of depth human silhouettes.

Depth video frames contain an unrestricted environment having a number of uncertain objects, obstacles, noisy background and freely movement of a human in the scenes. Therefore, to remove noisy effects in raw depth images, we propose a floor removal technique based on a least squares method [19–21] for estimating the angle and center point of...