Pattern Recognition Letters 48 (2014) 70–80

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Review Article

Human activity recognition from 3D data: A review ☆

J.K. Aggarwal, Lu Xia *

The University of Texas at Austin, Austin, TX 78705, USA

ARTICLE INFO

Article history:
Available online 4 May 2014

Keywords:
Computer vision
Human activity recognition
3D data
Depth image

ABSTRACT

Human activity recognition has been an important area of computer vision research since the 1980s. Various approaches have been proposed with a great portion of them addressing this issue via conventional cameras. The past decade has witnessed a rapid development of 3D data acquisition techniques. This paper summarizes the major techniques in human activity recognition from 3D data with a focus on techniques that use depth data. Broad categories of algorithms are identified based upon the use of different features. The pros and cons of the algorithms in each category are analyzed and the possible direction of future research is indicated.

1. Introduction

This paper is a tribute to the work of Dr. **Maria Petrou** on 3D data. She pioneered many techniques and methodologies dealing with 3D data which have benefited researchers as well as industry. She developed techniques for generating a 3D map from photometric stereo [3]; she proposed two methodologies of characterizing 3D textures: one based on gradient vectors and one on generalized co-occurrence matrices [45]; she built algorithms to reconstruct 3D horizons from 3D seismic datasets [8] and to inference the shape of a block of granite from cameras placed at 90 degrees to each other [30]. Her passing on October 15, 2012, just as the new advanced 3D sensors were becoming available, was a great loss. As far as we know, she did not present work on activity recognition from 3D data. Even though, she made important contributions to the computer vision community and inspired the later works on 3D which includes the application on activity recognition [17]. In this paper, we are going to give a review of the recent works on activity recognition using 3D data.

Recognizing human activity is one of the important areas of computer vision research today. The goal of human activity recognition is to automatically detect and analyze human activities from the information acquired from sensors, e.g. a sequence of images, either captured by RGB cameras, range sensors, or other sensing modalities. Its applications include surveillance, video analysis, robotics and a variety of systems that involve interactions between persons and electronic devices. The development of human activity recognition from depth sensors began in the early 1980s. Past research has mainly focused on learning and recognizing activities from video sequences taken by visible-light cameras. Those works were summarized at different depths from different perspectives in several survey papers [1,83]. The major issue with visible-light videos is that capturing articulated human motion from monocular video sensors results in a considerable loss of information. This limits the performance of video-based human activity recognition. Despite the efforts of the past decades, recognizing human activities from videos is still a challenging task.

After the recent release of cost-effective depth sensors, we see another growth of research on 3D data. We divide the methods to obtain 3D data from the past 20 years into three categories. One way is by using marker-based motion capture systems such as MoCap.[1] The second way is from stereo: capture 2D image sequences from multiple views to reconstruct 3D information [3]. The third way is to use range sensors. The development of range cameras has progressed rapidly over the past few years. Recently, the advent of depth cameras at relatively inexpensive costs and smaller sizes gives us easy access to the 3D data at a higher frame rate resolution, leading to the emergence of many new works on action recognition from 3D data. We will discuss the state-of-the-art algorithms on human activity recognition using 3D data in each category in this article with a focus on recent developments on depth data.

Depending on the environments, human activity may have different forms ranging from simple actions to complex activities. They can be conceptually categorized into four categories [1]: atomic actions, activities that contain a sequence of different actions, interactions that include person–object interactions and

☆ This paper has been recommended for acceptance by Edwin Hancock.

* Corresponding author. Fax: +1 512 471 5532.
E-mail address: xialu@utexas.com (L. Xia).

[1] http://mocap.cs.cmu.edu/.

http://dx.doi.org/10.1016/j.patrec.2014.04.011
0167-8655/

person-to-person interactions, and group activities. Research on atomic action recognition from 3D data has developed for many years, while complex activities and interactions have been studied recently after the easy access of 3D data became available. The research on group activities using 3D data is limited, either due to the difficulty of obtaining the data or limitation of the sensors.

Here we enumerate four major challenges to vision based human action recognition. The first is low level challenges [93,13]. Occlusions, cluttered background, shadows, and varying illumination conditions can produce difficulties for motion segmentation and alter the way actions are perceived. This is a major difficulty of activity recognition from RGB videos. The introduction of 3D data largely alleviates the low-level difficulties by providing the structure information of the scene. The second challenge is view changes [63,93,37,92,68]. The same actions can generate a different "appearance" from different perspectives. Solving this issue with a traditional RGB camera is achieved by introducing multiple synchronized cameras, which is not an easy task for some applications. But this is not a serious problem for recognition algorithms using a 3D motion capture system. For recognition from range images, this problem is partially alleviated because the appearance from a slightly rotated view can be inferred from the depth data. Even though, the problem is not totally solved because the range image only provides information on one side of the object in view, nothing is known about the other side. If skeletal joint information can be inferred accurately using a single depth camera, a view-invariant recognition algorithm may be constructed from the skeletal joint information. The third challenge is scale variance, which can result from a subject appearing at different distances to the camera or different subjects of different body sizes. In RGB videos, this can be solved by extracting features at multiple scales [14]. In depth videos, this can be easily adjusted because the true 3D dimension of the subject is known [78,97]. The fourth challenge is intra-class variability and inter-class similarity of actions [64]. Individuals can perform an action in different directions with different characteristics of body part movements, and two actions may be only distinguished by very subtle spatio-temporal details. This still remains a hard problem for algorithms using various types of data.

The objective of this article is to provide an overview of the state-of-the-art methodologies on human-activity recognition using 3D data. We discuss various techniques to acquire 3D spatio-temporal data, and summarize recognition methodologies using each type of data source. In particular, we will talk about 3D from stereo, 3D from motion capture system and 3D from depth sensors. Most of the recent works on human activity recognition reside in the third category, and some of the techniques used are adapted from previous works in the first two categories. We, therefore, put our focus on the third category in this survey. Although there are also works on 3D shape from shading [66], 3D from focus/defocus [41,27], 3D from texture [53], 3D from motion [16] and so on, they are not covered in this survey. These problems are usually ill-posed and the solutions are not unique [65]. Even with parameters such as the light source, the surface reflectance and the camera, it is still hard to get rid of some ambiguities. Hence, the algorithms are mostly proposed to solve the 3D shape of static objects. To the best of our knowledge, there is no work on activity recognition that uses shading, defocus texture, or motion to acquire 3D spatio-temporal information.

Recently, [12] wrote a survey summarizing the human motion analysis algorithms using depth imagery. Their paper summarizes depth sensing techniques, preprocessing of depth images, pose recognition, and action recognition from depth video, datasets, and libraries. However, the activities recognition methodologies were only coarsely categorized and introduced. This article will follow a different framework, and we will show a deeper analysis of the activity recognition algorithms. Especially, we will give an inside look at the features that were proposed in different scenarios and cover more recent publications. The activities discussed in this review range from atomic actions to complex activities, which include whole body activities, upper-body gestures, hand gestures, person–object interactions and person-to-person interactions. Static gestures, gait, and pose recognition are not in the scope of the paper.

2. 3D from stereo

Stereo, the reconstruction of three-dimensional shapes from two or more intensity images is a classic research problem in computer vision [54,67]. Early range sensors are expensive and cumbersome; the low-cost digital stereo camera has therefore generated interest in vision-based systems. A stereo camera is equipped with two or more lenses with a separate image sensor or film frame for each lens. This allows the camera to simulate human binocular vision, and therefore gives it the ability to generate three-dimensional images. By comparing the two images, the relative depth information can be obtained, in the form of disparities, which are inversely proportional to the differences in distance to the objects. The detail of stereo matching and computing the disparity/depth from the matching is not discussed in this paper, please refer to [25] for details.

Stereo vision is highly important in fields such as robotics, and it also has application in entertainment, information transfer, and automated systems. For instance, Dr. Maria Petrou developed a photometric stereo technique, which uses a fixed digital camera and three lights to illuminate the object from different angles. All the data are combined into one 3D image by analyzing the shadows and highlights. This technique has been used to find flaws in industrial surfaces and to capture a 3D map of a human face [3]. The techniques of recovering and analyzing depth images made a great contribution to the computer vision community. Kanade et al. developed a video-rate stereo machine that generated a dense depth map at the video rate [43] and demonstrate its application in merging real and virtual worlds in real time.

As mentioned above, we draw the ideas of acquisition of 3D information from Petrou, Kanade, and other researchers. Further, researchers combine the 3D with other information to analyze sequences of images. In particular, researchers further explored its application on tracking [20,59,69,105], pose recognition [18,58,75,32] and human activity recognition from stereo images or sequences [33,85]. The commonly used activity recognition schemes fall into three categories.

The first category is model based approaches which fit a known parametric 3D model, usually a kinematic model, to the stereo videos and estimate motion parameters of the 3D model. For instance, [84] co-register a 3D body model to the stereo images and extract joints locations from the body model and use joint-angle features for action recognition. The fitted 3D body model is less ambiguous than silhouettes, however, recovering the parameters and pose of the model is usually a difficult problem without the help of landmarks.

The second category is holistic approaches, which directly model actions using image formation, silhouettes, or optical flow [92]. Action is recognized by comparing the observations with learned templates of the same type. This usually requires that test and learned templates are obtained from similar configurations. [61] represent human actions as short sequences of atomic body poses, and the body poses are represented by a set of 3D silhouettes seen from multiple viewpoints. In this approach, no explicit 3D poses or body models are used, and individual body parts are not identified. [2] encode action using the Cartesian component