Information Fusion 40 (2018) 45–56

Contents lists available at ScienceDirect

Information Fusion

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

Multi-view stacking for activity recognition with sound and accelerometer data

Enrique Garcia-Ceja*, Carlos E. Galván-Tejada, Ramon Brena

Tecnologico de Monterrey, Av. Eugenio Garza Sada 2501 Sur, Monterrey, NL 64849, Mexico

ARTICLE INFO

Article history:
Received 5 December 2016
Revised 5 April 2017
Accepted 11 June 2017
Available online 12 June 2017

Keywords:
Multi-view learning
Stacked generalization
Activity recognition
Accelerometer
Sound

ABSTRACT

Many Ambient Intelligence (AmI) systems rely on automatic human activity recognition for getting crucial context information, so that they can provide personalized services based on the current users' state. Activity recognition provides core functionality to many types of systems including: Ambient Assisted Living, fitness trackers, behavior monitoring, security, and so on. The advent of wearable devices along with their diverse set of embedded sensors opens new opportunities for ubiquitous context sensing. Recently, wearable devices such as smartphones and smart-watches have been used for activity recognition and monitoring. Most of the previous works use inertial sensors (accelerometers, gyroscopes) for activity recognition and combine them using an aggregation approach, i.e., extract features from each sensor and aggregate them to build the final classification model. This is not optimal since each sensor data source has its own statistical properties. In this work, we propose the use of a multi-view stacking method to fuse the data from heterogeneous types of sensors for activity recognition. Specifically, we used sound and accelerometer data collected with a smartphone and a wrist-band while performing home task activities. The proposed method is based on multi-view learning and stacked generalization, and consists of training a model for each of the sensor views and combining them with stacking. Our experimental results showed that the multi-view stacking method outperformed the aggregation approach in terms of accuracy, recall and specificity.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Ambient Intelligence (AmI) [1] is an emerging discipline that brings intelligence to our everyday environments by adapting those environments to our needs [1], making them aware of the context [2]. It builds upon advances in sensors, pervasive computing, and artificial intelligence. AmI technologies should be sensitive, responsive, adaptive, transparent, ubiquitous, and intelligent. In an AmI environment, devices are expected to work collectively by sharing information and using the history of past events. The AmI vision puts lighting, sound, vision, domestic appliances, and personal health care products to cooperate seamlessly in order to help the user [3].

Several of these systems are based on Human Activity Recognition (HAR) since knowing the current activity is of great importance to understand the users' context. Following Dey's notion of context [4], the user's task is one of its key elements. This is why HAR research has received great interest recently [5–13]. Being able to detect the activities being performed by an individual can provide valuable information in the process of understanding the context and situation in a given environment, so it is of great interest because of the wide range of possible applications such as in medicine, Ambient-Assisted Living [14], sports, marketing [15], surveillance [16], etc.

Data from several sources is collected and then analyzed to extract useful context information in many HAR applications, and some works have explored the combination of different types of sensors for activity recognition [5–7]. However, most of them use an *aggregation* approach, i.e., extract features from each sensor and aggregate them to train a predictive model. Aggregation is not optimal since each sensors' data have their own statistical properties [17] and combining them in the same model can confound them. In this paper, we present a method based on multi-view learning and stacked generalization for fusing audio and accelerometer sensor data for human activity recognition using wearable devices. We treat each sensor's data as different *views* and then, they are combined using stacked generalization [18]. The proposed approach is flexible since it does not rely on a specific classification model and is efficient in terms of memory and computation since it only requires to train a classifier for each sensor type and an ex-

* Corresponding author.
E-mail addresses: e.g.mx@ieee.org (E. Garcia-Ceja), ericgalvan@uaz.edu.mx (C.E. Galván-Tejada), ramon.brena@itesm.mx (R. Brena).

http://dx.doi.org/10.1016/j.inffus.2017.06.004
1566-2535/© 2017 Elsevier B.V. All rights reserved.

tra meta-classifier. We evaluated the multi-view stacking approach for home tasks activity recognition using sound and accelerometer data collected with a smartphone and a wrist-band. Furthermore, we evaluated the proposed method with other three multi modal HAR datasets.

This document is organized as follows: Section 2 presents an overview of activity recognition and sensor fusion methods that have been used. This Section also presents the background on multi-view learning and stacked generalization. In Section 3 we explain how stacked generalization is used in the context of multi-view learning for activity recognition. Section 4 details the accelerometer/audio data collection process. In Section 5 we explain the feature extraction process for the audio and accelerometer data. Next, in Section 6 the experiments and results are presented and finally in Section 7 we draw the conclusions.

2. Related work and background

There are two main types of sensors that have been used for Human Activity Recognition: *external sensors* and *wearable sensors*. External sensors are installed in the environment and may not have direct physical contact with the user. Examples of such sensors are: video cameras, microphones, motion sensors, depth cameras like the Microsoft Kinect, RFID tags, switches, etc. On the other hand, wearable sensors are carried by the user or are embedded in devices such as smartphones, smartwatches and fitness bracelets. Examples of wearable sensors are: accelerometers, gyroscopes, magnetometers, to name a few.

Automatic activity recognition systems have been successfully developed using external sensors such as video cameras [19–21] and color-depth cameras [8,9]. With the recent advent of smartphones and wearable devices such as smart-watches and fitness bands, it is now possible to collect data from their different sensors without the need of a fixed infrastructure. Recently, the sensors embedded in those type of devices have been used for human activity recognition given their flexibility, ubiquity and unobtrusiveness. Often, inertial sensors (accelerometers and gyroscopes) are used for HAR tasks, albeit other types of sensors like microphones, light sensors, temperature, heart rate, etc. are also embedded in those type of wearable devices. The use of wearable sensors [22] has gained a lot of attention because they have several advantages; in particular, the recognition can be performed in any place unlike video cameras in which it is restricted to a specific area. Another problem of external sensors is that in environments with multiple residents it becomes difficult to detect which person activated a specific sensor. This is not a problem for wearable sensors since they are personal. Yet other problems of external sensors are related to privacy issues, because the user does not decide if s/he is going to be monitored.

A common recent trend is to use smartphones for HAR since they are immensely popular and they already have several types of embedded sensors. Another advantage is that all the processing can be performed inside the phone so there is no need to carry another processing unit. One of the first works to perform all the recognition inside a phone was the one of Brezmes et al. [23]. There are also other works that have used smartphones for activity recognition [10–13]. Given the advantages of wearable sensors, in this work we focus on this type of systems. Specifically, we used a smartphone and a wrist-band to perform the recognition.

2.1. Sensor fusion in activity recognition

With the increasing miniaturization of sensors, it is now common to find many types of them in our environment, especially in wearable devices. Given their ubiquity and sensing capabilities, a wide range of physical phenomena can be measured, thus, generating large quantities of diverse data types. These data can be used to extract contextual information from the environment allowing the realization of reactive systems based on the current inferred state. Combining the diverse sources of data in an intelligent manner in order to generate knowledge and extract useful information has been an active research area [24]. In multimedia analysis, it is common to have different sensing modalities such as video, audio, text, WWW resources, etc. and the fusion of the multiple sources can increase the accuracy of the system [25]. In medical image analysis, the fusion of different imaging modalities (MRI, Ultrasound, CT, PET, SPECT) can produce improved results [26]. For human activity recognition, the most common approach is to use inertial sensors (accelerometers, gyroscopes, tilt switches, etc.) due to their flexibility and infrastructureless capabilities; however, external sensors (cameras, switches, motion sensors, RFID) are also used, mainly in smart environments.

In the work of Tolstikov et al. [27] Dynamic Bayesian Networks and Dempster-Shafer theory are used to fuse sensor data in order to recognize seven activities of daily living using 14 binary sensors placed around a house. Their results suggest that both methods are similar in terms of performance. Amoretti et al. [28] also used Bayesian Networks to monitor activities for ambient assisted living in a smart environment. In wearable sensor settings, the most common approach of sensor fusion is *aggregation*, i.e., concatenate the extracted features from all sensors and train a single classification model with them. Shoaib et al. [5] explored the use of smartphones' accelerometers and gyroscopes tested individually and in combination for activity recognition. They used an aggregation approach and found that the combination of both sensors improved the overall performance when the individual performances are not very high. In a later work [6] of the same authors (Shoaib et al.) they combined smartphone and wrist-worn inertial sensors by aggregation and–again– obtaining better results when fusing both devices. Dernbach et al. [7] also conducted experiments with accelerometer and gyroscope sensors and also concluded that combining information from both sensors (by aggregation) increased the system performance by 10–12% accuracy. In a similar work, Hayashi et al. [29] also used an aggregation of accelerometer and sound data to classify daily activities achieving better results when using both sources of information.

Even though it has been shown that combining multiple sources of data can increase the system accuracy, *aggregation* is not optimal since each sensor has its own statistical properties [17] which may require a different treatment. Zhu & Sheng [30] used a multi-sensor fusion scheme for combining sensors attached to the foot and waist. They trained two different neural networks for each of the sensors to recognize coarse-grained activities. Then, the outputs of the neural networks are fused using manually defined rules. These rules dictate if the fine-grained activity classification should be performed by heuristic discrimination or a Hidden Markov Model. By training two different neural networks they are able to preserve the statistical properties of each sensor unit; however, the manual definition of rules becomes hard when increasing the number of activities, thus, limiting the scalability of the approach. Another sensor fusion method was proposed by Banos et al. [31] whose aim is to be robust against hardware failures. Their proposed classifier is trained in a hierarchical fashion by first generating $m \times n$ binary classifiers where m is the number of sensors and n the number of classes, i.e., n binary classifiers for each sensor. Then, these classifiers are weighted and aggregated with a particular function for each sensor which corresponds to the second level classifiers. Finally, the decisions of the second level classifiers are weighted and combined to produce the final prediction. This method proved to be very robust in the presence of sensor failures by combining information of the

Download English Version:

https://daneshyari.com/en/article/4969119

Download Persian Version:

https://daneshyari.com/article/4969119

Daneshyari.com