Human Action Classification in Basketball: A Single Inertial Sensor Based Framework

Xiangyi Meng[1], Rui Xu[1], Xuantong Chen[1], Lingxiang Zheng[1(✉)], Ao Peng[1], Hai Lu[1], Haibin Shi[1], Biyu Tang[1], and Huiru Zheng[2]

[1] School of Information Science and Engineering, Xiamen University, Xiamen, China
lxzheng@xmu.edu.cn
[2] School of Computing and Mathematics, University of Ulster, Newtownabbey, UK
h.zheng@ulster.ac.uk

Abstract. Human Action Recognition is becoming more and more important in many fields, especially in sports. However, conventional algorithm are almost camera-based methods, which make it cumbersome and expensive. As the wearable inertial sensor has developed a lot, in this paper, we present a novel human action classification algorithm using in basketball, based on a single inertial sensor, which is a application of multi-label classification. We performed experiment on real world datasets. The AUPRC, AUROC and confusion matrix of our results demonstrated that our novel basketball motion recognizer have a great performance.

Keywords: Basketball motion · Human action recognition
Single inertial sensor · Feature extraction · Support vector machine
Multi-label classification

1 Introduction

In recent years, human action recognition (HAR) becomes more and more useful in many ares, including some human-computer interaction (HCI) applications like somatic game, human health monitoring, robotics [2,3,14]. Formally, the aim of HAR is to automatically detecting, analyzing and recording human actions from information and data obtained from many sources both on-line and off-line, for example, wearable inertial sensors, annotated video segments, etc. [3,11]. Consequently, in terms of sensors type used in HAR applications, there are two principal method for HAR: vision-based HAR and inertial-based HAR [3].

An ideal way to recognize human actions is to use the vision information. Shuiwang et al. proposed a 3D convolutional neural network for human action recognition on RGB video data, which can handle 3D inputs [5]. With the development of the techniques used for depth extraction from video [7], many deep learning based HAR approach using depth video data are proposed [8,12,13].

J. C. Hung et al. (Eds.): FC 2017, LNEE 464, pp. 152–161, 2018.
https://doi.org/10.1007/978-981-10-7398-4_16

However, those methods just perform well on the existing datasets, not showing their strengths on real situations. Preliminarily, to deploy those approaches on real world stages must satisfy those prerequisites: (1) Applicable cameras to get a large amount of undimmed video segments; (2) Proper and interruption-free place to setup the cameras; (3) Powerful CPU/GPUs to run deep learning algorithm efficiently.

Moreover, consider the case of a non-professional basketball player. Peter, a skillful programmer, is an amateur basketball player who proposed to develop an application helping himself to train shooting skill. Intuitively, he made the computer capable to capture his action each time he shot. As mentioned above, the best solution seemed to be a video sensor HAR system. However, it turned out inconvenient to setup the cameras and supporting devices before his shooting training. Such case is common in many situations of HAR with the challenges of occlusion, camera position, computational complexity, etc. [3], though, it performs well on large scale datasets. These limitations constrained the applications of HAR, especially where too many noises exist. To address such problems, empirical researches studied the HAR based on wearable inertial sensors with accelerometers and gyroscopes, which is convenient enough for individuals to use in their daily routines.

An effective way to analyze the basketball action is to dig out the characteristics in the data obtained from the hand used during the action, which helps coaches and athletes to evaluate their performance better and optimize the training projects. This problem can be regarded as a multi-class classification problem with some inevitable issues [1], as shown below:

Intraclass Variability. This is the first challenge of HAR that a well-performed HAR framework must be robust to the intraclass variability. Those variabilities are common because the same action might be performed differently by different individuals. For example, Stephen and Shawn Marion (an 20-plus-point scorer in the early 2000s NBA) have definitely different shooting postures, whereas they can both get many scores during a match.

Interclass Similarity. Plus there is a inverse challenge, namely, interclass similarity, meaning that different actions are fundamentally different but they have similar numerical characteristics. For example, considering two common actions: **shooting** and **high lobbing pass**, both of them need players to lift their hand and force the ball out of their hands, consequently returning similar sensor data, respectively.

The NULL class problem. Typically, only limited parts of motion types are manually classified and can be recognized by the HAR system. It's an intuition that given this imbalance of relevant versus irrelevant data, activities of interest can easily be confused with activities that have similar patterns but that are irrelevant to the application in question–the so called NULL class. Of course, in some certain HAR applications, such as basketball shooting, golf, etc., the NULL class problem is not particularly serious, given that the types of the motion is not too complex.

To address these problems, we designed a new motion classifier deployed on basketball motion recognition. Being different with other motion recognition frameworks that need many sensors providing several kinds of data like RGB, depth, inertial data, etc., our recognizer is just based on the single inertial sensor, which returns accelerometers and gyroscopes data, attached on the user's wrist of the shooting hand. Then we proposed a novel feature extraction formula to.

The rest of this paper is organized as follows. Section 2 introduces the related work of the sports motion recognition using wearable sensors. The construction and processing of our dataset are stated in Sect. 5.1. Then we defined our problem mathematically in Sect. 3. Section 4 includes the feature extraction method and basketball motion classification method. Then we deployed our methods and the results are shown and analyzed in Sect. 5. Finally, Sect. 6 summarizes our paper.

2 Related Works

In recent years, wearable devices, for example, smart-watch, smart bracelet, etc., have gained unprecedented development. Due to their portability and low power consumption, wearable devices play an important role in the area of activity monitoring, performance evaluation and feedback providing. Andrea et al. presented an comprehensive survey on human activity recognition using body-worn inertial sensors [1]. In this survey, the authors limn the background and some state-of-the-art HAR frameworks, which can be characterized as an process, named Activity Recognition Chain (ARC), giving researchers a clear and understandable tutorial of HAR. The ARC is a process combining the method of signal processing, pattern recognition and machine learning techniques, which receives the raw data returned from the sensors as an input, and responds an output carrying the classification result of the action corresponding to the raw data. That is a common framework of HAR. Following, some relevant researches with in the field of sports motion recognition using wearable sensors are introduced.

Technical statistics are important in sport competitions. However, it is time-consuming and boring to do that manually. Now, with the help of the wearable devices, the technical actions can be recognized and recorded automatically. Taking rugby as an example, Kelly et al. addressed the problem of the automatic recognition of the tackles and collisions in rugby using a GPS receiver and an accelerometer placed between the shoulder blades overlying the upper thoracic spine of each player. In detail, they applied support vector machine (SVM) and hidden conditional random field (HCRF) to identify those actions above, resulting into an excellent performance, where the recall and the precision were 93.3% and 95.8%, respectively.

As for swim, Bächlin et al. designed a wearable assistant, named SwimMaster. What make the SwimMaster helpful is its real-time performance evaluation system using the swimming parameters extracted from the data obtained from the sensors embedded in the assistant.

Le et al. studied the basketball activity recognition problem using wearable inertial measurement units (IMU) [9]. However, being different from our work, they

deployed 5 IMUs on the body (two are on the foots, two are on the legs and the remaining one is on the back of the user), which is obviously uncomfortable and will constrain the actions of the user when playing basketball, intuitively.

3 Problem Definition

This section gives information about the basic idea of identifying the motions of playing basketball and the main problem we are facing.

3.1 Prerequisite

In order to effectively analyze performance of a basketball player, a precise identification of the entire motion is essential, namely, to distinguish shooting from other types of motion. Now the only thing we have is a sensor placed on the twist of the habitual basketball shooting hand to record the tri-axial accelerated and angular velocity of the chip, also, those of the wrist.

3.2 Basketball Shooting Recognition

The task of basketball motion recognition is to correctly recognize the type of a given motion from a number of motions belonging to various kinds. In a basketball match, shooting, pass and dribble are three kinds of motions of great importance. So our paper mainly investigated the classification of the three types of motions. In order to facilitate the description of our approach, here we defined the problem more mathematically.

Considering one of a whole process of the human motion, let's denote it as $\mathcal{S}_i = \left(\mathbf{d}_i^1, \mathbf{d}_i^2, \ldots, \mathbf{d}_i^{|S_i|}\right)^{\mathrm{T}}$, representing the i-th motion in a test case. Formally, $\mathbf{d}_i^j = (a_x, a_y, a_z, g_x, g_y, g_z)$ consists of the data from the accelerometer and gyroscope. In addition, $\mathcal{S}$ represents the set of the motions in a test case. The task of basketball shooting recognition is to find a judging function $f : \mathcal{S} \to \mathcal{Y}$, where $\mathcal{Y} = \{y_1, y_2, \ldots, y_{|\mathcal{S}|}\}$, $y_i \in 0, 1, 2$ is a set of judging results for whether the motions in $\mathcal{S}$ is basketball shooting, pass or dribble.

Intuitively, this can be formalized as a multi-label classification problem. To build the classifier, for each motion, a feature vector $\mathbf{x}_i$ should be derived using the information of $\mathcal{S}_i$. Then, our task reduces to build a model to estimate the probability $P(y_i|\mathbf{x}_i)$. However, it is challenging to accurately define and compute $\mathbf{x}_i$ given that each motion $\mathcal{S}_i$ has its own length and the law of the data is hard to mine. In the next section, a novel feature extraction method were explained, then we deployed a multi-label SVM to classify the motions.

4 Methods

This section illustrates the methods we use to overcome the complicated analyzing process and to obtain credible identification of basketball motion. To give a clear illustration about our approach, here we give the flowchart of our proposed method, shown in Fig. 1.

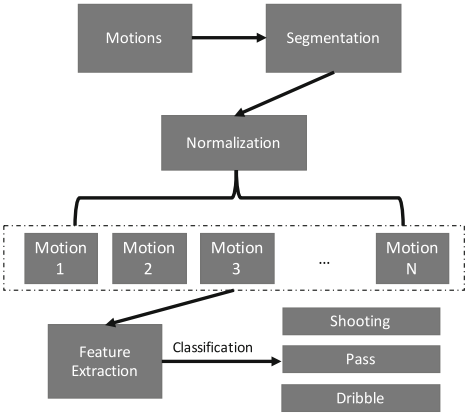

Fig. 1. This figure shows the flowchart of our basketball motion recognition method.

4.1 Feature Extraction

For each motion $\mathcal{S}_i$, it can be denoted as a matrix,

$$\mathcal{S}_i = \begin{pmatrix} d_i^1 \\ d_i^2 \\ \vdots \\ d_i^{|\mathcal{S}_i|} \end{pmatrix} = \begin{pmatrix} a_x^1 & a_y^1 & a_z^1 & g_x^1 & g_y^1 & g_z^1 \\ a_x^2 & a_y^2 & a_z^2 & g_x^2 & g_y^2 & g_z^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_x^{|\mathcal{S}_i|} & a_y^{|\mathcal{S}_i|} & a_z^{|\mathcal{S}_i|} & g_x^{|\mathcal{S}_i|} & g_y^{|\mathcal{S}_i|} & g_z^{|\mathcal{S}_i|} \end{pmatrix} \tag{1}$$

where each row in the matrix $\mathcal{S}_i$ represents a sample of the sensor data. Notice that different matrix $\mathcal{S}_i$ has its own size of row since motions are different in various of aspects. So we need to scale them into the same size which is set as the maximum size of each sample $max\{|\mathcal{S}_i|\}, i = 1, 2, \ldots, |\mathcal{S}|$. Therefore, before computing the feature vector, an extrapolation must be implemented to finish this preliminary.

Subsequently, we need to calculate the average vector of each row of the matrix, denoted as

$$\overline{d_i} = \sum_{j=1}^{|\mathcal{S}_i|} d_i^j = (\overline{a_x}\ \overline{a_y}\ \overline{a_z}\ \overline{g_x}\ \overline{g_y}\ \overline{g_z})\,. \tag{2}$$

The average vector of the sensor data characterizes the comprehensive numerical feature, whereas we need to ensure our feature vector $\mathbf{x}_i$ is able to carry all of the information useful to describe an motion. An intuition is to calculate the

distances between each row vector and the average vector, as defined below:

$$\mathbf{x}_i = \begin{pmatrix} dist(d_i^1, \overline{d}_i, \Sigma) \\ dist(d_i^2, \overline{d}_i, \Sigma) \\ \vdots \\ dist(d_i^{|\mathcal{S}_i|}, \overline{d}_i, \Sigma) \end{pmatrix} \tag{3}$$

where $dist(\mathbf{x}, \boldsymbol{\mu}, \Sigma)$ is the distance function. In this paper, we chose Mahalanobis distance as our metric, denoted as

$$dist(\mathbf{x}, \boldsymbol{\mu}, \Sigma) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})} \tag{4}$$

where $\mathbf{x} = (x_1, ..., x_n) \in \mathbb{R}^n$, $\boldsymbol{\mu} = (\mu_1, ..., \mu_n) \in \mathbb{R}^n$ and

$$\Sigma = \begin{pmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & \cdots & E[(X_1 - \mu_1)(X_n - \mu_n)] \\ \vdots & \ddots & \vdots \\ E[(X_n - \mu_n)(X_1 - \mu_1)] & \cdots & E[(X_n - \mu_n)(X_n - \mu_n)] \end{pmatrix}.$$

Finally, the feature vector $\mathbf{x}_i$ of an entire specific motion $\mathcal{S}_i$ has been computed. Considering that the dimensionality of the feature vector $\mathbf{x}_i$ is too high, before deploying the classification algorithm, we performed Principal Component Analysis (PCA) to reduce the dimensionality.

4.2 Multi-label Support Vector Machine

Generally, there are two strategy, **one-vs-all** anb **one-vs-one**, for multi-label classification. In this paper, considering the potential inter-class similarities, we chose the **one-vs-one** strategy. Based on the selected strategy, we built a scalable linear support vector machine to classify the motions.

5 Experiments and Results

5.1 Dataset

In this section, we will introduce the composition of our dataset. Briefly, our dataset used in this paper is a combination of two sources: one is obtained from our own data collection process, the other is the UTD Multimodal Human Action Dataset (UTD-MHAD) [4].

Data Collection Process. In order to get the inertial data during an basketball motion, we deployed an *Arduino 101* board worn on the wrist of the experiment candidates.

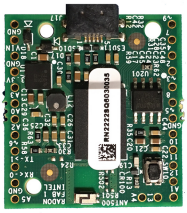

Fig. 2. The mini Arduino 101 board we used in this paper.

Arduino 101 board has a 6-axis accelerometer/gyro and onboard Bluetooth LE (BLE) capabilities, shown in Fig. 2. Then we developed an android application to receive and store the data transferred from the board via BLE, moreover, this application can control when the board should stop collecting data and send the data back. Finally, we got 157 shooting motions, 70 pass motions and 80 dribble motions.

UTD-MHAD. UTD Multi-modal Human Action Dataset is a part of a research on human action recognition in The University of Texas at Dallas, US. It is a fusion of depth and inertial sensor data. In its collecting process, only one Kinect camera and inertial sensor were used. There are 27 kinds of human actions in UTD-MHAD, including basketball shooting, tennis serve, pickup and throw, etc. In this paper, we only used the single inertial sensor to recognize the basketball motion, consequently we only picked up the inertial data of basketball shooting, which is a subset of the UTD-MHAD, to supplement our experiments. The subset consists of 32 packages of sensor data of basketball shooting generated from 8 experiment candidates.

In order to accurately verify the performance of our proposed model, here we design a leave-one-out cross validation (LOOCV) for the training and prediction process. LOOCV is a strategy to increase the accuracy of the evaluation of the classifier, for which every sample in the sample set have the chance to be the test sample, and the other samples are regarded as the training set. LOOCV was selected as the method to construct the training and test set because the training set generated by the LOOCV accommodates almost all of the samples, which makes sure that the training set is quite similar to the original distribution of the samples.

Based on the dataset introduced in Sect. 5.1, we implemented our algorithms described in Sect. 4 using *Python* and its popular scientific calculation package *SciPy* [6] and a machine learning package *scikit-learn* [10].

Here we evaluated the classification performances using the area under the curve for Receiver Operating Characteristic Curves (AUROC) and Precision Recall Curves (AUPRC).

5.2 Results

The average AUROC and AUPRC of our experiment using the method described in Sect. 4 are summarized in Table 1. Moreover, as the Confusion Matrix showed in Fig. 3, the dribble recognition shows the best performance because of the dribble motion is significantly different from the other two motions. In addition, our classifier may sometimes confused shooting motions and pass motions for the reason that these two motions are somewhere similar with each other. That intuitively makes sense.

Table 1. The performances of the basketball motion classification using features we extracted.

Class	AUROC	AUPRC
Shooting	0.731	0.756
Pass	0.804	0.839
Dribble	0.967	0.982

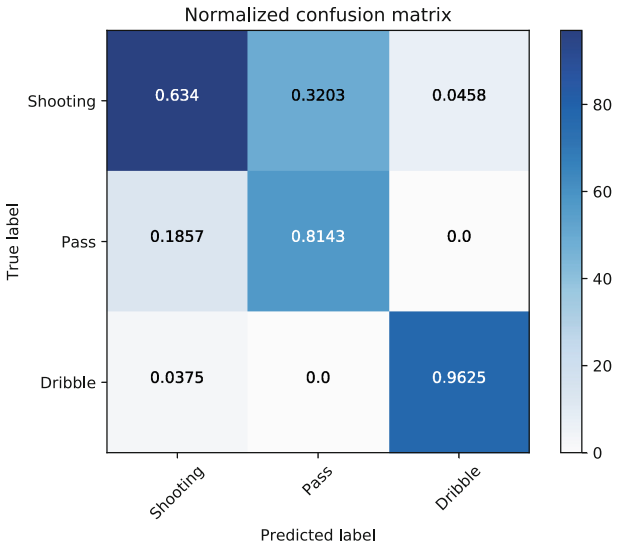

Fig. 3. The normalized confusion matrix of the multi-label SVM.

6 Conclusion

In this paper, we presented a novel human action recognition framework for basketball. Being different with conventional study on the human action recognition, our proposed framework is based on single inertial sensor, which is convenient to be deployed in real world situation and moreover, energy-saving. Our framework consists mainly of two parts: feature extraction and classification. Naive binary classifier is not enough to solve our multi-motion recognition task. To address problems in our task, we deployed a multi-label SVM, which fits the basketball motion recognition problem well. Moreover, the experiment based on the real world datasets demonstrated that our framework performed well in the task of basketball motion recognition.

However, some problems are still worth being studied. First, in our current framework, we need the whole sensor data, from the beginning to the end, to identify whether a motion is a motion. However, this makes it hard to deploy our framework in real time applications. Second, there still exists some singular

motion posture, which is totally different with common style, but they are still effective. Current classifier has none knowledge about that "new" motion, which will lead to ridiculous mistakes. Moreover, the computation of the multi-label SVM is time-consuming and under the current calculation ability of portable devices, we need to send data to high performance servers to perform our algorithm. Unfortunately, the time wasted during the communication further decrease the possibility of deploying our framework in real time applications. Thus those aspects are what we should next focus on.

Acknowledgments. This work is supported by Student's Platform for Innovation and Entrepreneurship Training Program, Xiamen University (2016Y1123), and 2016 Google Student Innovation Project (64008066).

References

1. Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput. Surv. (CSUR) **46**(3), 33 (2014)
2. Chen, C., Jafari, R., Kehtarnavaz, N.: Improving human action recognition using fusion of depth camera and inertial sensors. IEEE Trans. Hum.-Mach. Syst. **45**(1), 51–61 (2015)
3. Chen, C., Jafari, R., Kehtarnavaz, N.: A survey of depth and inertial sensor fusion for human action recognition. Multimed. Tools Appl. **76**(3), 4405–4425 (2015)
4. Chen, C., Jafari, R., Kehtarnavaz, N.: UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 168–172. IEEE (2015)
5. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
6. Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python (2001). http://www.scipy.org/
7. Karsch, K., Liu, C., Kang, S.B.: Depth extraction from video using non-parametric sampling. In: European Conference on Computer Vision, pp. 775–788. Springer (2012)
8. Kuo, W.Y., Kuo, C.H., Sun, S.W., Chang, P.C., Chen, Y.T., Cheng, W.H.: Machine learning-based behavior recognition system for a basketball player using multiple kinect cameras. In: 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), p. 1. IEEE (2016)
9. Nguyen, L.N.N., Rodríguez-Martín, D., Català, A., Pérez-López, C., Samà, A., Cavallaro, A.: Basketball activity recognition using wearable inertial measurement units. In: Proceedings of the XVI International Conference on Human Computer Interaction, p. 60. ACM (2015)
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
11. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)

12. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Commun. ACM **56**(1), 116–124 (2013)
13. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp. 568–576 (2014)
14. Xu, Y., Shen, Z., Zhang, X., Gao, Y., Deng, S., Wang, Y., Fan, Y., Chang, E.I., et al.: Learning multi-level features for sensor-based human action recognition. arXiv preprint arXiv:1611.07143 (2016)