From Kinect skeleton data to hand gesture recognition with radar

Jiayi Li1, Aman Shrestha2, Julien Le Kernec1,2*, Francesco Fioranelli2

1School of Information and Electronics, Univ. of Electronic Science and Technology of China, Chengdu, China
2Communication, Sensing and Imaging Group, School of Engineering, University of Glasgow, Glasgow, UK
*Julien.le Kernec@glasgow.ac.uk; Julien.le Kernec@uestc.edu.cn

Keywords: MICRO-DOPPLER, RADAR, HUMAN ACTIVITY RECOGNITION, MACHINE LEARNING, CLASSIFICATION

Abstract

In an era where man-machine interaction increasingly uses remote sensing, gesture recognition through use of the micro-Doppler effect is an emerging application which has attracted great interest. It is a sensible solution and here we show its potential for detecting aperiodic human movements. In this paper, we classify ten hand gestures with a set of handcrafted features using simulated micro-Doppler signatures generated from Kinect skeleton data. Data augmentation in the form of synthetic minority oversampling technique has been applied to create synthetic samples and classified with the Support Vector Machine and K-nearest neighbour classifier with classification rate of 71.1% and 51% achieved.

Finally, using weights generated by an action pair based one vs one classification layer improves classification accuracy by 24.7% and 28.4%.

1. Introduction

With the rapid development of IoT and smart homes (i.e. smart light or air conditioner remotely controlled by humans), the demand for systems able to recognize human motions and react accordingly has arisen. Waving hands or drawing a circle are different physical movements, and these 2 actions can be treated as 2 instructions to control different appliances. Gesture classification helps computers recognise human motions, to provide another form of man-machine interface.

Much research focuses on Human Activity Recognition (HAR) from optical sensors such as visible light camera [1], [2]. Features extracted from silhouettes and spatial-temporal domains are widely used in HAR [3]. Another video representation is action graph using Gaussian mixture models with salient postures via unsupervised learning [4]. There is a wealth of work in camera/video HAR [5]. Even though radar presents more technical challenges than video for HAR, it is a preferred sensing modality because it does not record plain images of people. It protects the user sense of privacy and there is currently no legal/ethical framework regulating assisted living technologies, so image rights might become a legal issue in the future [6]. Compared with optical sensors, radar can be operated in dark environments which means that artificial light is unnecessary for HAR [7] [8] [9]. The motion of limbs will lead to frequency modulations in spectrograms around the main Doppler component [10]. Classification in radar has thus been using micro-Doppler (mDs) signatures from which features are extracted to recognise activities [5] [11] [12][13].

To date, 3 approaches have been used to simulate human mD signatures: kinematic modelling [10], video motion capture (MOCAP)[14] and Kinect [15] [16]. Previous work mainly focused on classifying periodic movements e.g. run, walk and crawl [17] [18] [19], and torso information can be extracted from the mean Doppler component for classification.

This paper reports on the work conducted in a final year project at the Glasgow College – UESTC in an undergraduate EEE degree programme in 2017-2018 looking at the issue of classifying aperiodic motions such as hand gestures.

2. Methodology

2.1. The Kinect sensors

The first Kinect was released in 2010 for interactive games [20]. In this project, the Kinect V2 was used and comprises a RGB camera and a IR camera capable of extracting the 3D skeleton data. Table 1 lists the main features of Kinect.

The advantages of Kinect lie in its low price and depth sensing capability (0.4m-8m). Moreover, a skeleton tracking toolbox is available on MATLAB [21] and other development platform. However, the IR camera is sensitive to external infrared source and cannot detect highly reflective objects [15].

Kinect sensor is of sufficient quality to simulate human mD signatures for classification purposes [15]. However, our preliminary tests have shown an instability in skeleton extraction that would require correcting the skeleton before being used for mD simulation as it would create unrealistic Doppler modulations because of the glitches as shown in Fig.1.

Table 1 Basic parameters of Kinect V2 sensor[20]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Kinect V2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth Sensing Technology</td>
<td>Time of flight</td>
</tr>
<tr>
<td>Colour Image Resolution</td>
<td>1920x1080 30fps (12fps low light)</td>
</tr>
<tr>
<td>IR Image Resolution</td>
<td>512x424 30fps</td>
</tr>
<tr>
<td>Depth Sensing Resolution</td>
<td>512x424 30fps</td>
</tr>
<tr>
<td>Field of View</td>
<td>&gt; 43° vertical, 70° horizontal</td>
</tr>
<tr>
<td>Depth Sensing Range</td>
<td>0.5-4.5m, up to 8m without skeletonising</td>
</tr>
<tr>
<td>Skeleton Tracking</td>
<td>Up to 6 subjects</td>
</tr>
<tr>
<td>(with full skeleton)</td>
<td>25 joints per skeleton</td>
</tr>
</tbody>
</table>

Fig.1: Walking human model (unit: meters) - left) normal plot (right) anomalous skeleton extraction
Because of time constraints for the completion of the project and the time it would take to work on designing a more robust skeleton extraction routine, it was decided to work from Kinect databases that had already been “cleaned up”.

2.2. MSR Action3D Dataset

The Kinect data required for simulation is from MSR Action3D Dataset [22] captured at 15 fps. It includes depth map sequences with 640x240 resolution and skeleton data in screen and 3D Cartesian coordinates. The database contains 20 action types performed 2-3 times by 10 subjects. Due to high-frequency vibrations and incorrect skeleton tracking issues in some recordings, we picked the most reliable records of 9 actions (Table 2) each having 20 samples. Only the skeleton data is used to construct the human model.

<table>
<thead>
<tr>
<th>No</th>
<th>Action description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High arm wave</td>
</tr>
<tr>
<td>2</td>
<td>Horizontal arm wave</td>
</tr>
<tr>
<td>3</td>
<td>Hammer</td>
</tr>
<tr>
<td>4</td>
<td>Hand catch</td>
</tr>
<tr>
<td>5</td>
<td>Forward punch</td>
</tr>
<tr>
<td>6</td>
<td>Draw X</td>
</tr>
<tr>
<td>7</td>
<td>Draw Stick</td>
</tr>
<tr>
<td>8</td>
<td>Draw circle</td>
</tr>
<tr>
<td>9</td>
<td>Hand clap</td>
</tr>
</tbody>
</table>

Table 2 List of human actions

2.1 Human model reconstruction

To construct the skeleton, 17 of the detected joints are included with the Boulic kinematic model [23] and the V. Chen’s mD modelling [10] (Fig. 2 left). The radar return from human motion was emulated with the radar cross section (RCS) of ellipsoids snapped on the skeleton (Fig. 2 right).

\[
\text{rect}(t) = \begin{cases} 1, & 0 \leq t \leq \Delta \\ 0, & \text{otherwise} \end{cases}
\]  

(2)

The body is composed of spheres and ellipsoids for the radar cross section (RCS) estimation. As spheres are particular ellipsoids, the RCS is calculated for both using (3) [24].

\[
RCS_{\text{ellip}} = \frac{\pi a^2 b^2 c^2}{(a^2 \sin^2 \theta + b^2 \sin^2 \theta + c^2 \cos^2 \theta)^2}
\]  

(3)

The spectrogram is obtained using the short-time Fourier transform of the return signals following the method in [10].

2.4. Simulated human micro-Doppler signatures and feature extraction

The 15fps Kinect dataset is too sparse to derive an accurate spectrogram [15], the skeleton data provided by [22] is interpolated to achieve a sampling rate of 2048 Hz.

A pulse-Doppler radar is simulated with a carrier frequency of 15 GHz. The radar is placed at the right-hand side (10m,0,0) of the subject. In other words, it is at 90 degrees from the direct line of sight of the Kinect camera.

<table>
<thead>
<tr>
<th>Features name and numbers</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entropy (#1)</td>
<td>The randomness of energy distribution of the spectrogram</td>
</tr>
<tr>
<td>Skewness (#1)</td>
<td>Asymmetry of energy distribution of the spectrogram</td>
</tr>
<tr>
<td>Centroid (mean &amp; variance)</td>
<td>Centre of mass of the spectrogram</td>
</tr>
<tr>
<td>Bandwidth (mean &amp; variance)</td>
<td>An estimation of energy distribution range around the centroid</td>
</tr>
<tr>
<td>SVD (mean &amp; variance of right and left vectors) (#12)</td>
<td>Decompose the energy contents into time and frequency domain.</td>
</tr>
<tr>
<td>Upper / Lower envelope (min, max &amp; standard deviation) (#2)</td>
<td>Statistic features of the upper and lower envelope of spectrogram</td>
</tr>
<tr>
<td>Torso (bandwidth &amp; average frequency)</td>
<td>A measure of features related to torso movement</td>
</tr>
<tr>
<td>Energy curve (mean, standard deviation &amp; integrated) (#3)</td>
<td>Energy content of the spectrogram</td>
</tr>
<tr>
<td>Time duration between peaks of the upper envelope (minimum, mean &amp; maximum) (#3)</td>
<td>The repetition of peaks of spectrograms in time domain</td>
</tr>
</tbody>
</table>

2.3. Radar returns

From Fig. 3, the radar return signal is a time-delayed version of the transmitted signal. Thus, for a sequence of narrow rectangular pulses with transmitted frequency $f_c$, a pulse width $\Delta t$, and a pulse repetition interval $\Delta T$, the radar return from a body part may be represented by (1), which is given by [10].

\[
s_p(t) = \sum_{k=1}^{n_p} \sigma_p(t) \text{rect}[t - k \Delta T \frac{R_p(t)}{c}] \exp[-j2\pi f_c \frac{2R_p(t)}{c}]
\]  

(1)

where $\sigma_p(t)$ is the RCS of the body part at time $t$, $n_p$ is the total number of pulses received, $R_p(t)$ is the distance from the radar to the body part at time $t$, and the rectangular function $\text{rect}$ is defined by (2).

![Fig. 2 left] 17 joints based human skeleton structure[10], right) Human model of waving hands action (X1 = 10m, Y1 = 0m, Z1 = 0m)

![Fig. 3] Spectrogram top) without MTI bottom) with MTI
Before extracting features from the spectrogram, a single delay line canceler is applied as a Moving Target Indicator (MTI) filter [24] to remove the torso contributions around the zero-Doppler line; as it provides little information and masks more salient mD signatures i.e. improves contrast (Fig. 4).

Features (Table 3) were extracted from spectrograms: entropy and skewness (information of the energy distribution), envelope and energy curve, Singular Value Decomposition (SVD - decomposing the energy contents into time and frequency domain), time durations between peaks, upper and lower Doppler bandwidth. Other parameters were tested as classically seen in periodical motion analysis [9]. However, after initial testing the features (Table 3) that are not greyed were abandoned as they were not salient for this classification task. The others bear a number indicating how many features are used e.g. SVD (#12) has 12 features for classification.

2.2 Data augmentation

The 20 samples of each action are too sparse for machine learning algorithms to generalize the models accurately. To augment the number of samples, synthetic minority over-sampling technique (SMOTE) [25] was applied to create more “synthetic” samples. It generates samples in “feature space” instead of “data space”. After choosing the 20 samples to 15 training and 5 test samples per action, SMOTE is used applied separately to the datasets with an oversampling rate of 500% and the number of nearest neighbours K = 3. We will explain why in the results part. The augmented dataset contains 75 training and 25 test samples as shown in Fig. 4.

3. Hand gesture classification results

Two classifiers SVM with the RBF kernel function [26][27] and KNN with K = 5 [28] are used throughout.

3.1 Classification of 9 actions (1 vs all)

The features (Table 3) are used as inputs to classifiers based on supervised machine learning. After running the proposed oversampling and classification for 100 times, the results are averaged and we generated the confusion matrices as shown in Fig. 5. The KNN accuracy is 72.1% and that of SVM is 51%. Both the overall accuracy and confusion matrix show that the classification accuracy is rather low. Actions are frequently misclassified as other actions. One oddity of this result is that KNN performs better than SVM. By looking more closely at the SMOTE algorithm, it relies on KNN to create “synthetic” samples as explained earlier. Instead of performing randomly assigning the samples between training and test data sets first and then using SMOTE, the separation between training and test sets was performed at first prior to data augmentation. The simulations were repeated 100 times and the results averaged using the method in section 2.2 for data augmentation. The KNN accuracy dropped to 27.4% and that of SVM to 31.7%; the confusion matrices are shown in Fig. 6.

Fig. 5: Confusion matrices for KNN (left) and SVM (right) classifiers for SMOTE augmentation and then random training and test sample separation.

The accuracy is as low as the original data without SMOTE augmentation which means that the synthetic samples are better representing reality. This method is therefore retained to pursue an optimization of the classification accuracy.

3.2 Classification of action pairs

Due to the relatively low classification accuracy aforementioned, action pairs classification is carried out. The aim is to simplify the feature space that currently has 22 features to classify 9 actions in a 1 vs all modality. We now switch to a 1 vs 1 classification approach to discover the salient features. The test classifies an action pair with only 1 feature involved. For each action pair, the 22 features are tested separately and the feature yielding the highest accuracy is retained as shown in Table 4. Using feature #13, actions 1 and 2 can be classified with an accuracy of 100%.

Table 4: The accuracy of classification of action 1 and action 2 by using a single feature by using KNN

Fig. 6: Confusion matrices for KNN (left) and SVM (right) classifiers for random training and test sample separation then SMOTE data augmentation.

Fig. 7: 1-feature best classification for one versus one classification.

3
A summary of the best classification accuracies of action pairs is shown in Fig. 7. Using the cumulative number of times, a feature was picked to best classify an action pair, a table of feature weights was defined for the 22 features (Table 5).

Table 5: feature weights based on the one versus one test

<table>
<thead>
<tr>
<th>Feature number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Feature number</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>Frequency</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

KNN classifier

<table>
<thead>
<tr>
<th>Feature number</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Feature number</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>Frequency</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

3.3. Classification of 1 vs all with weighted features

To optimize the performance of classification, weighted feature classification approach is applied. The weights \(w\) (Table 5) are applied to the features. After 100 iterations, the KNN accuracy is 52.1% and that of SVM is 60%. The confusion matrix is shown in Fig. 8.

Fig. 8: Confusion matrices for the KNN (left) and SVM (right) classifiers for our proposed weighted features method.

The KNN and SVM classifiers show similar accuracy results. The confusion between action 6 (draw x) and 8 (draw circle) can be explained by the similarity of the movements when projected in the radar radial direction. Actions 1, 3 and 4 are easily misclassified as almost all actions.

One noticeable result is that actions 2,5,7 and 9 have a high accuracy and are not misclassified with one another. On that basis, we reran the classification for these 4 actions only and got an accuracy of 94.2% with KNN and 97.7% with SVM.

4. Analysis and Discussions

4.1. Data augmentation with SMOTE

In the classification result, the SMOTE algorithm significantly increased the classification accuracy of KNN classifier (raised from 31.1% to 72.1%) also with a smaller increment in the accuracy of SVM classifier (raised from 26.7% to 51.0%). When using SMOTE algorithm before dividing training and testing samples, it had a positive impact on the result. However, it means that the synthetic samples are creating a biased dataset and do not represent the original samples accurately. The data augmentation introduced bias. When looking at the way SMOTE generates samples, neighbours are the key to generate more samples. It is also easy to notice that KNN classifier is using neighbours to classify samples. Thus, SMOTE is biased towards the neighbour-based classifier, thus explaining the higher accuracy and therefore the result cannot be trusted.

To avoid this bias, using SMOTE algorithm after dividing the training and test samples was adopted (section 2.2). When SMOTE algorithm generates the two groups of data independently, the neighbours of training data may not necessarily be the neighbours of test data and provides a more realistic “synthetic” dataset.

The classification result also shows the fact that the SMOTE after dividing samples is a reliable data augmentation technique. The result does not vary too much (KNN 27.4% SVM 31.7%) from the original dataset before augmentation.

4.2. One vs one compared to one vs all

The difference in classification results of action pairs and all 9 actions are noticeable. The accuracy of one versus one with 1 feature has a very high accuracy greater than 80% for most pairs. However, the accuracy of all actions classification is lower than 40%.

Such a gap between the 2 classifications is mainly caused by the complexity of the model (too many features were not salient) and high similarity between actions from the radar perspective. With the proposed weighted feature approach, a simpler model was reached and the accuracy improved by about 30% for both KNN and SVM classifiers. By setting the weights of the irrelevant features to 0 and the most salient features to larger weights (Table 5), the overall accuracy almost increased from 27.4% to 52.1% for KNN classifier and 31.7% to 60.0% for SVM classifier. Therefore, in feature-based classification, the old adage less is more applies.

4.3. Similarity of actions

The similarity of actions is also an important factor that will lead to the degradation in classification accuracy. Similar actions are highly likely to be misclassified with one another which makes the classification more challenging as observed in Fig. 9 and 10 for confusing actions and distinguishable ones respectively.

Radar can only sense the ranging information from a target. So, a lot of spatial information in other directions are lost when using a radar to detect the movements. Actions which are not alike in three dimensions may have a greater similarity when projected on the radial direction. For example, the confusion between action 6 (draw x) and 8 (draw circle) can be explained by the similarity of the physical movements. The projection of these two actions on the direction of radar is almost the same.

This problem may not be solved with one radar to detect the movement. However, by placing different radar sensors at different angles as shown in [29] may help to solve this problem as well as studying the aspect angle dependence of the hand gesture as studied in [30]. Information lost in one direction can be captured in another direction.

4.4. Comparison to the state of the art

In investigating of using mD radar to classify human actions, most researchers are focused on classifying locomotion (e.g. walking, running, leaping, crawling and creeping) [15] [19]. These motions contain a lot of periodic information and torso oscillation. The classifications of these motions are using the features related to the bandwidth of the torso oscillation and average torso velocity.
This project is to classifying actions (e.g. high arm wave, hand clap, draw stick) that have almost no information in these two features mentioned above and are aperiodic.

The difficulty of classifying these actions by using mD radar is that the discriminating features are different for different action pairs, instead of a few features being sufficient to classify all actions which makes the one versus all classification more challenging.

5. Conclusion
A Kinect-based mD radar simulator was developed and tested in this paper. Classification experiments were carried out to recognise human hand gestures based on simulated spectrograms.

For the data preparation, the limited records for each action in the MSR dataset led us to the SMOTE data augmentation technique in order to obtain a sufficient number of samples for machine learning. We also highlighted a problem with the use of such techniques that should be considered when using it.

Our proposed weighted feature technique allowed an increase of about 30% in classification accuracy. The weighted feature technique removes irrelevant features and increases the weight of salient features. Furthermore, after removing the most confusing actions, the accuracy of 4 actions increased to 97.7% with SVM showing the relevance of this technique.

Future work will look at carrying out live experiments for validation with ultra-wideband software defined radar [7] [9] [13] [31] and [32].

6. Acknowledgements
A. Shrestha is supported for his PhD by the UK Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Award to the School of Engineering.

The authors acknowledge support from the UK EPSRC through grant EP/R041679/1 INSHEP.

References
Fig. 9: Spectrograms of actions 3, 4, 6 and 8 that confused with each other by the classifiers.

Fig. 10: Spectrograms of actions 2, 5, 7 and 9 that are differentiated with high accuracy by the classifiers.