



## Fusing Thermopile Infrared Sensor Data for Single Component Activity Recognition within a Smart Environment

Burns, M., Morrow, P.J., Nugent, C.D., & McClean, S. I. (2019). Fusing Thermopile Infrared Sensor Data for Single Component Activity Recognition within a Smart Environment. *Journal of Sensor and Actuator Networks*, 8(1), 1-16. Article 10. <https://doi.org/10.3390/jsan8010010>

[Link to publication record in Ulster University Research Portal](#)

### Published in:

Journal of Sensor and Actuator Networks

### Publication Status:

Published (in print/issue): 18/01/2019

### DOI:

[10.3390/jsan8010010](https://doi.org/10.3390/jsan8010010)

### Document Version

Author Accepted version

### General rights

The copyright and moral rights to the output are retained by the output author(s), unless otherwise stated by the document licence.

Unless otherwise stated, users are permitted to download a copy of the output for personal study or non-commercial research and are permitted to freely distribute the URL of the output. They are not permitted to alter, reproduce, distribute or make any commercial use of the output without obtaining the permission of the author(s).

If the document is licenced under Creative Commons, the rights of users of the documents can be found at <https://creativecommons.org/share-your-work/licenses/>.

### Take down policy

The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact [pure-support@ulster.ac.uk](mailto:pure-support@ulster.ac.uk)



1 Article

# 2 Fusing Thermopile Infrared Sensor Data for Single 3 Component Activity Recognition within a Smart 4 Environment

5 Matthew Burns, Philip Morrow, Chris Nugent and Sally McClean

6 School of Computing, Ulster University; {burns-m19, pj.morrow, cd.nugent, si.mcclean}@ulster.ac.uk

7 Received: date; Accepted: date; Published: date

8 **Abstract:** To provide accurate activity recognition within a smart environment, visible spectrum  
9 cameras can be used as data capture devices in solution applications. Privacy, however, is a  
10 significant concern with regards to monitoring in a smart environment, particularly with visible  
11 spectrum cameras. Their use may therefore may not be ideal. The need for accurate activity  
12 recognition is still required and so an unobtrusive approach is addressed in this research  
13 highlighting the use of a Thermopile Infrared Sensor as the sole means of data collection. Image  
14 frames of the monitored scene are acquired from a Thermopile Infrared Sensor highlighting only  
15 sources of heat, for example, a person. The recorded frames feature no discernable characteristics of  
16 people hence privacy concerns can successfully be alleviated. To demonstrate how Thermopile  
17 Infrared Sensors can be used for this task, an experiment has been conducted to capture almost 600  
18 thermal frames of a person performing four single component activities. The person's position  
19 within a room along with the action being performed are used to appropriately predict the activity.  
20 The results demonstrate that high accuracy levels of 91.47% for activity recognition can be obtained  
21 when only using Thermopile Infrared Sensors.

22 **Keywords:** Thermopile; Infrared; Sensors; Activity Recognition; Image Processing; Sensor Fusion;  
23 Activities of Daily Living; Computer Vision; Smart Environments.  
24

---

## 25 1. Introduction

26 It has been predicted that the world's population is expected to reach as high as 8.6 billion by  
27 2030 [1]. It is also predicted that the number of people requiring 24/7 monitoring and care, whether  
28 due to a disability or an age-related issue, will also increase. Due to the detrimental psychological  
29 effects of moving into a nursing home and that almost 90% of over 65s that prefer living at home [2],  
30 it is preferable to facilitate someone remaining at home for as long as possible. The term, *aging in*  
31 *place*, refers to this concept and can be defined as the ability, irrespective of age or salary, to  
32 independently and safely live at home [3].

33 Activities of Daily Living (ADLs) embody the day to day actions and activities that we perform  
34 independently for our own self-care. The items that fall under this category are activities such as  
35 feeding ourselves, bathing, grooming and dressing [4]. The analysis of the completion of such  
36 activities can benefit the monitoring the health and wellbeing of residents through the detection of  
37 medical issues, lifestyle changes in addition to age-related diseases [5]. Monitoring the actions and  
38 ADLs of a person in their own home provides the ability to understand their routine which  
39 subsequently allows a better appreciation of what aid is required to benefit the person the most. This  
40 understanding can help to facilitate the delivery of the care essential for allowing a person to remain  
41 at home.

42 The monitoring of a home environment can be made possible through the deployment of sensors  
43 that will continuously collect relevant data and the subsequent processing of the data. Many

44 approaches exist which can be deployed for recognising ADLs based on sensor data. In [6] an  
45 approach to ADL recognition for streaming sensor data within a smart home was proposed. Several  
46 ADLs were covered in this approach, including grooming, sleeping, eating, cleaning, washing and  
47 preparing meals. Sensor data was streamed and segmented into individual parts, with the intention  
48 that each segment represented the sensor events that had been triggered for a single activity. This  
49 segmentation was carried out using a sliding window where the segments were used to populate  
50 rows of training data which the chosen machine learning model, a Support Vector Machine (SVM),  
51 processed. The data generated from each separate sensor was separated so that each segment would  
52 ideally represent one activity due to the existing knowledge of the beginning and ending of sensor  
53 events triggered by the activities. This training data consisted of the activity, times for the start, end  
54 and duration of the activity and each individual sensor tag which also indicated whether the sensor  
55 had fired. The primary reason for using two continuous sliding windows was to compare the  
56 probability of correctness for each window's activity prediction. This then highlighted whether the  
57 probability trend was going up or down. To evaluate the results of the study, both five and ten-fold  
58 cross validation were implemented, producing an overall accuracy of 66%, with each activity causing  
59 a significantly visible variance amongst their individual accuracies. Activities that underachieved  
60 with regards to performance and accuracy were found to have had less training data, showing the  
61 necessity for a sufficiently large dataset.

62 Three popular categories of devices used to capture data are wearable devices, visible spectrum  
63 cameras and thermal infrared cameras. For example, in [7] wearable sensors are used to detect ADLs,  
64 where Inertial Measurement Units (IMUs) were used to collect and process data from actions such as  
65 sitting down, standing up, reaching high and low, turning and walking. A mock up apartment was  
66 set up to facilitate the participants' completion of a cleaning task. The task was laid out in a manner  
67 that the participants needed to perform the previously stated actions to complete it. For example,  
68 objects were placed at various heights to force the participant to reach out at different heights and  
69 armchairs were placed within the environment to prompt sitting down and standing up actions. This  
70 allowed the system to attempt to predict the action at any given time. Each participant was required  
71 to complete the task in three, four and five-minute durations. Five randomly chosen five-minute trials  
72 were used for the training of the recognition algorithms, with all three and four-minute trials used to  
73 test the algorithms. Participants wore a motion capture suit made up of seventeen IMUs where the  
74 acceleration, angular velocity and 3D orientation of each IMU was captured at a frequency of 60Hz.  
75 During the task, kinematic peaks identified an activity where the activity was segmented by taking  
76 the maximum/minimum to the left/right of the peaks to estimate the activity's duration. Kinematic  
77 and angular data was extracted from the relevant body parts for each of the actions and the activities  
78 were detected and classified using the sensor signals at an accuracy of approximately 90%. The  
79 average median time difference between the manual and sensor segmentation was approximately  
80 0.35 seconds. While promising accuracies were achieved in this study, wearable devices are not  
81 preferred as alternatives to video sensors due to required maintenance and having to wear electronic  
82 equipment [8].

83 The use of computer vision / image processing technologies for activity recognition may provide  
84 a more non-invasive approach, since there is no requirement for the use of any wearable technology.  
85 The study in [9] shows that there are clear benefits to being able to incorporate image processing  
86 techniques into the task of recognising activities. Such benefits include the use of segmentation for  
87 detecting human movements or the various motion tracking algorithms facilitated by computer  
88 vision-based approaches. RGB-D cameras have also been used where depth information has been  
89 incorporated with the image data [10]. Here, the camera was positioned on the ceiling with the  
90 intention of predicting a performed action and, as a result, detect abnormal behavior. This work  
91 considered each ADL to be predicted as a set of sub-activities or actions. A set of Hidden Markov  
92 Models (HMMs) were employed and trained using the Baum-Welch algorithm [11] to be able to  
93 accurately detect any significant changes in states. The position of a person's head and hands in 3D  
94 space were detected and recorded for the input for the models. The three HMMs involved were  
95 configured to receive input from the head, the hands and the head and hands together, respectively.

96 The five activities to be predicted were daily kitchen activities: *making coffee, taking the kettle, making*  
97 *tea or taking sugar, opening the fridge and other*. Here, *other* encompasses all other kitchen related  
98 activities. Each model individually recognised the sequence of activities and predicted the overall  
99 activity accordingly. The model that produced the highest probability for its prediction was chosen.  
100 The classification results of the experiment were produced from a test where 80 trials were used to  
101 train the model with a further 20 trials being used for testing. The model tailored for the head  
102 obtained an average f1-score of 0.80, with the model created for only the hands generated an average  
103 f1-score of 0.46. Finally, the model that made use of both the head and hands data obtained a 0.76  
104 average f1-score. Visible spectrum cameras, however, can give rise to a level of discomfort within the  
105 home space, due to their obtrusive nature. This can bring about a lack of natural behavior from the  
106 home's inhabitants. While they allow for the collection of useful and rich data, these security and  
107 privacy concerns have previously been highlighted by those who are subject to monitoring [3]. Such  
108 concerns can act as a roadblock for the successful production of activity recognition systems built  
109 with obtrusive elements. These concerns require addressing.

110 An unobtrusive alternative to cameras that operate on the visible spectrum, are devices that  
111 make use of thermal imagery or data. In [12] a thermal sensor is used to classify various postures and  
112 detect the presence of a person. A method of background subtraction was implemented where a  
113 threshold value was used to remove any pixels that were not associated with the person in the  
114 environment. A class referring to the data collected when nobody was present in the environment  
115 was used to calculate this threshold. The features that were extracted from the data included the  
116 difference between both the threshold and the highest detected temperature, as well as the number  
117 of pixels with values larger than the threshold. The total, standard deviation and average gray levels  
118 from the pixels that made up the person were also calculated. The classification of the data was  
119 conducted by decision tree models built using Weka's J48 supervised learning algorithm. The  
120 training dataset was generated from data collected over three days and, based on 10-fold cross-  
121 validation, the model achieved 90.67% and 99.57% for pose and presence recognition, respectively.  
122 The two testing datasets were generated from data on two separate days where the first test dataset  
123 produced 75.95% and 99.94% for pose and presence recognition, respectively. Accuracies of 60.06%  
124 for pose recognition and 91.65% for presence detection were achieved with the second test dataset. It  
125 was found that the results for the second set of test data suffered as the data was captured at a higher  
126 room temperature. It was concluded that a greater variety in the training data with regards to a larger  
127 range of ambient temperatures was required to improve the overall levels of performance.

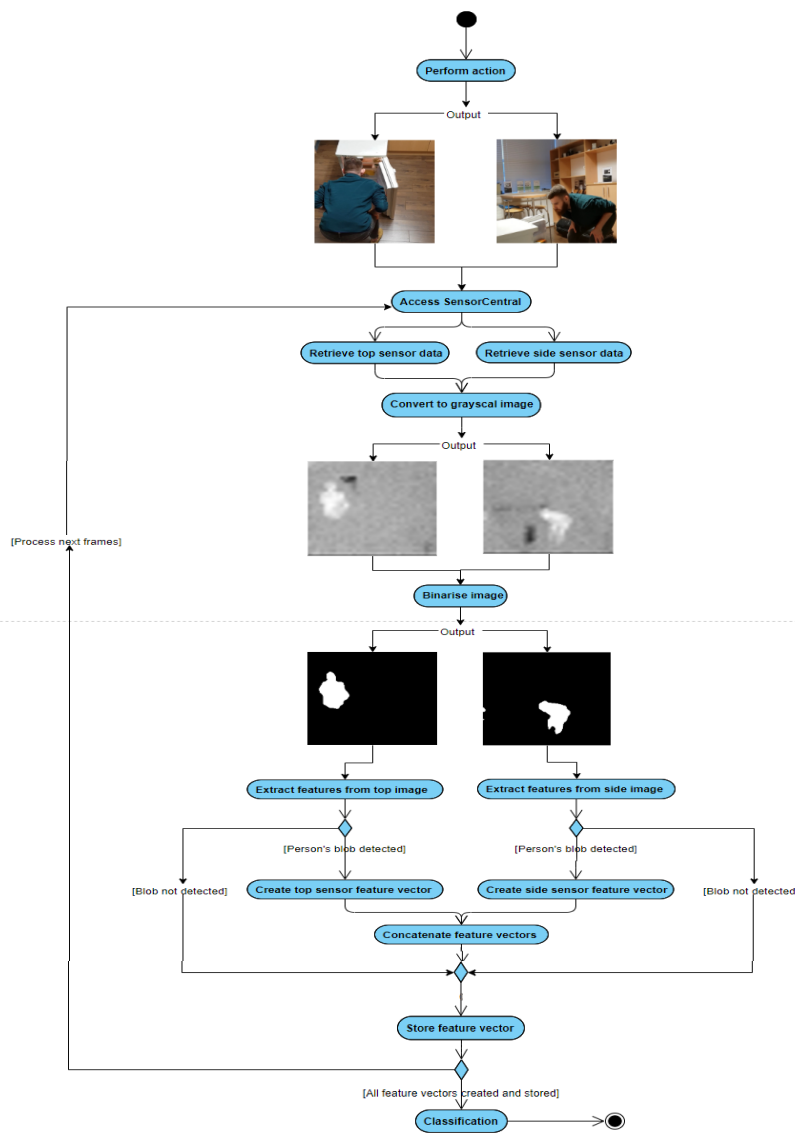
128 The Thermopile Infrared Sensor (TIS) [13] can be used to detect sources of heat, for example, a  
129 person. The collected data can then be output as a grayscale image. The image produced shows only  
130 areas of heat using a range of the pixels with the highest gray levels, with the lower grey level pixels  
131 signifying cooler areas. Intricate features of heat sources cannot be distinguished due to this lack of  
132 detail and resolution in the images and therefore, no discernable characteristics of people are able to  
133 be captured. In the work proposed in this paper we have used two TIS devices, situated to capture  
134 from two perpendicular planes. One of the devices was positioned on the ceiling of the environment  
135 and one on a tripod, surveying a side on view. The captured frames of the space are analysed to  
136 attempt to predict the activities being performed by the person in the room at any given time. This  
137 analysis process involves predicting the action of the person in each frame, using a collection of  
138 training data. The prediction is used along with the person's proximity to known objects in the room,  
139 such as the fridge or a table, to infer the likely activity.

140 This work aims to recognise single component activities including *opening/closing the fridge, using*  
141 *the fridge, using the coffee cupboard and sitting at the table*. These activities were chosen as they are  
142 common sub-activities of ADLs such as making a coffee or a meal. This allowed us to investigate  
143 whether the TISs would eventually be able to be used for such multiple component activities. This  
144 aim is to be fulfilled whilst sufficiently addressing any privacy concerns with regards to the capturing  
145 of images within the home. The advantageous factor of image processing techniques is intended to  
146 be retained in order to produce an accurate and unobtrusive activity recognition approach.

147 The remainder of this paper is structured as follows: Section 2 provides details of the platform  
 148 and methodology for activity recognition, using only the TIS. Section 3 outlines the single component  
 149 activity recognition experiment which was conducted, and Section 4 presents the results of the  
 150 experiment. The evaluation of the results, discussion and conclusions are presented in Section 5,  
 151 together with details of potential future work.

152 **2. Materials and Methods**

153 The research in this study has been carried out in the smart kitchen in Ulster University [14].  
 154 This environment is equipped with numerous sensors; including two 32x31 TISs which are located  
 155 on the ceiling and in the corner of the room. For this work we are only making use of only the TISs.  
 156 The two TISs are set up as sources for the *SensorCentral* sensor data platform [15]. The sensor data is  
 157 then provided by the *SensorCentral* sensor data platform in JSON format. An overview of the initial  
 158 stages of the implemented method is depicted in Figure 1, where the sensors have captured a person  
 159 bending at the fridge.



160  
 161

Figure 1. Overview of the initial stages of the method.

162 The fundamental functionality of this single component activity recognition approach is to  
 163 retrieve thermal frames from two sensors of the same type and extract and fuse relevant features to  
 164 predict the single component activity being performed within each frame. Upon determination of the  
 165 action being performed within the frame, the object that the person is nearest to is calculated. This  
 166 process can be viewed in the pseudo code in Figure 2.

```

SET nearestObjectDistanceXPlane TO 0
SET nearestObjectDistanceYPlane TO 0
FOR each frame pair
  FOR each object
    FOR each proximity point
      IF distance between BLOB's X centroid value and proximity point's X value < X plane threshold AND
distance between BLOB's Y centroid value and proximity point's Y value < Y plane threshold
        IF distance between BLOB's X centroid value and proximity point's X value <
nearestObjectDistanceXPlane AND distance between BLOB's Y centroid value and proximity point's Y
value < nearestObjectDistanceYPlane
          SET nearestObjectDistanceXPlane TO distance between BLOB's X centroid value and
proximity point's X value
          SET nearestObjectDistanceYPlane to distance between BLOB's X centroid value and
proximity point's X value
          SET nearestObject to object
        ENDIF
      ELSE
        SET nearestObject to NONE
      ENDIF
    ENDFOR
  ENDFOR
ENDFOR
ENDFOR
  
```

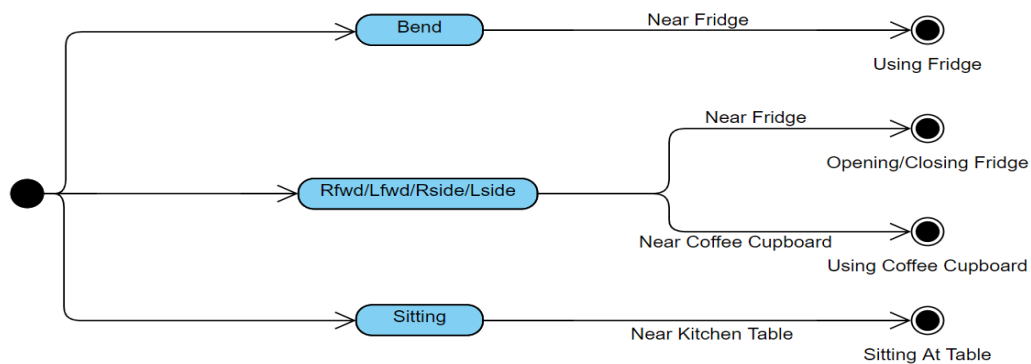
167  
 168

**Figure 2.** Pseudo for the process of calculating the nearest object.

169 Once it is determined if the person is close to an object in the frame and if so, what the object is,  
 170 it is used alongside the action prediction to infer the activity being performed within the frame. An  
 171 overview of this final aspect of the method can be viewed in Figure 3.

172 The first step in the process is to retrieve the thermal frames from *SensorCentral*, which acts as  
 173 the middleware for the devices and the developed system. The raw data captured by the TIS is  
 174 packaged in JSON format and consists of the frame data, timestamp and the sensor ID. The JSON  
 175 formatted frame data from both TIS devices is retrieved and used to fill a 32x32 matrix. For  
 176 convenience, the image is then resized to a 256x256 image. The TISs are, however, 32x31 sensors and  
 177 so this 32<sup>nd</sup> row is simply a black line of pixels which when the image is resized to 256x256, makes up  
 178 the bottom seven rows. These rows are removed, resulting in a 256x249 image. Once the frames from  
 179 both sensors are established, they are binarised using Otsu's automatic threshold method [16]. This  
 180 allows the person's shape to be analysed and features extracted to train the chosen machine learning  
 181 model. Frames from both TISs are captured at the same time and upon retrieval of a pair of these

182 frames, their timestamps are compared to ensure the frames were captured at the same instant and  
 183 not seconds or more apart.



184  
 185

**Figure 3.** Overview of the activity inference process.

186 The Binary Large Object (BLOB) depicting the person is found using the conditions that the  
 187 BLOB’s area is within pre-set parameters (chosen empirically), as well as it not having a similar  
 188 centroid position as the known objects within the room i.e. the fridge, coffee cupboard and the kitchen  
 189 table. Fourteen features are collected and extracted from both the shape of the person’s BLOB and the  
 190 pixels that make up their BLOB. The fourteen features from each frame in the pair are then combined  
 191 to form a twenty-eight-element feature vector. The same features are extracted from each of the  
 192 sensors. The features extracted from a sensor, along with brief descriptions, are detailed in Table 1.

193 Since the temperature of the person may fluctuate, causing a change in pixel grey levels, features  
 194 that target the person’s BLOB pixel values could not be used on their own. The standard deviation  
 195 and variance of the grey levels are still selected as features as they can still be somewhat useful in  
 196 differentiating between the person’s actions. It is, however, important to identify features that are  
 197 invariant to temperature change. Performing different actions causes the shape of the person’s BLOB  
 198 to noticeably change and so features that describe this shape are invaluable. The eccentricity of the  
 199 shape helps handle the changes in the shape’s elongation and so can help with detecting if the  
 200 person’s arms are being held out.

201 The convex area, equivalent diameter, solidity and the extent also aid in describing the shape of  
 202 the person’s BLOB. This is due to the large changes that occur to the width and height of the BLOB’s  
 203 shape during action transitions, but also the change in the area of the containing box or polygon when  
 204 the person, for example, bends, sits or just stands with their arms down. The ratio between the major  
 205 and minor axis also helps with such descriptions, where the choice to use the ratio between these  
 206 values was made to create a more variable feature, making it an easier task to separate actions.

207 These features help to differentiate between completely different actions, but it is the orientation  
 208 feature that is vital to determine the difference between more similarly shaped actions such as, for  
 209 example, facing a certain direction and holding the left arm out to the side and then holding the right  
 210 arm to the side but facing the opposite direction. Knowing the coordinates of the bounding box  
 211 encapsulating the BLOB also helps in differentiating between actions, most notably, whether it is the  
 212 right arm or left arm that is being extended. The features on their own describe specific attributes of  
 213 the BLOB but it is their combination that helps achieve the highest possible recognition rate.

214

215

**Table 1.** Features collected from each of the two TIS devices

Feature	Description
Eccentricity	The ratio of the distance between the foci of the shape’s ellipse and its major axis length
Major and minor axis ratio (Pixels)	Ratio between the length of the major axis of the ellipse and the length of the minor axis of the ellipse
Standard Deviation	Standard Deviation of the pixel grey levels within the detected BLOB
Variance	Variance of pixel grey levels within the detected BLOB
Bounding Box corner coordinates	The coordinates of each of the four corners making up the bounding box of the BLOB i.e. the smallest rectangle that can contain the BLOB.
Orientation (Degrees)	Angle between the <i>x</i> -axis and the major axis of the ellipse. The value is in degrees, ranging from -90 degrees to 90 degrees
Convex area	Number of pixels in the convex hull. This is the smallest convex polygon that can contain the region
Equivalent diameter (Pixels)	Diameter of a circle with the same area as the region
Solidity	Proportion of the pixels in the convex hull that are also in the region
Extent	Ratio of pixels in the region to pixels in the total bounding box (smallest rectangle containing the region)
Moment of the shape	Returns the central sample moment of the pixel grey levels that make up the shape

216

Once the features are calculated for a frame, the feature vector is stored. This is repeated until each of the frames retrieved from *SensorCentral* have been analysed and processed. The action being performed in each frame is manually labelled to provide ground truth data. The training dataset is made up of 3538 feature vectors which provided sufficient examples of each action. Examples of the actions targeted for prediction are shown below, in Table 2.

221

Several machine learning algorithms were tried and tested to evaluate which achieved the highest accuracy of activity classification. While the Support Vector Machine has a tendency to over fit, it was tested on the training data as it makes use of what is known as a kernel trick. This technique is effective at defining clearer differences between the classes, making the process of distinguishing between them, a much simpler one. This, however, requires an appropriate kernel function to be chosen. A decision tree was used as it requires little intervention for any data preparation as any missing data wouldn’t cause the data to split to allow the tree to be built. The random forest machine learning algorithm was also tested as it reduces the overfitting that can be caused by simple decision trees as well as bringing about less variance through its use of multiple trees.

230

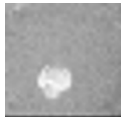
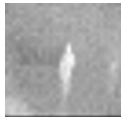
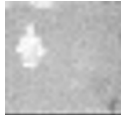











The primary advantage to employing a random forest model for this study is its effectiveness to estimate missing data. This is a scenario that is possible, as a frame retrieved from one of the two sensors may be unusable, leaving half of the feature vector empty. This may happen due to the accidental merging of the person’s BLOB with another object’s BLOB or due to a sudden spike of noise injected into the frame. Using 10-fold cross validation, the random forest model achieved the best accuracy score on the training set and so was used to recognise the single-component activities performed in the experiment.

236



237

**Table 2.** Thermal frame examples from the ceiling and side sensors

Action	Ceiling Sensor	Side Sensor
ArmsDown		
Bend		
Lfwd (Left Arm Forward)		
Rfwd (Right Arm Forward)		
Lside (Left Arm Extended to the Side)		
Rside (Right Arm Extended to the Side)		
Sitting		

238

239

240

241

242

243

244

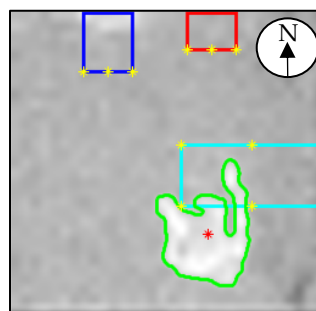
245

246

The locations of known objects within the space are also provided. These objects include the fridge, coffee cupboard and kitchen table. These objects are given what will be referred to as *proximity points*. The fridge and coffee cupboard have three proximity points each, located at their front left and right corners, and the middle of their south sides. The kitchen table has six proximity points positioned at its four corners and the middle of its north and south sides. These proximity points are plotted as yellow asterisks in Figure 4 which shows the view of the ceiling TIS where the person is sitting at the kitchen table (the cyan coloured rectangle). The dark blue rectangle represents the fridge, with the red rectangle representing the coffee cupboard. A compass has been annotated for reference.

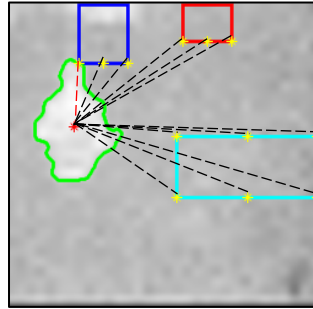
247

248



**Figure 4.** A person sitting at the kitchen table, as seen by the ceiling TIS

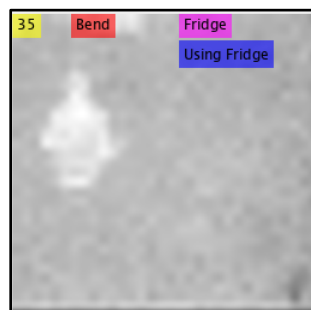
249 The information obtained from these objects was used to determine if the person was close to  
250 any of them by measuring the distance between the person's centroid and each object's proximity  
251 points. A diagram depicting this is shown in Figure 5 where the dashed line coloured red signifies  
252 the shortest distance between the person's centroid and a proximity point. As this proximity point  
253 belongs to the fridge, the person is predicted as being closest to the fridge.



254

255 **Figure 5.** Depiction of the distance measurement between the person's centroid and each object's proximity  
256 points

257 The label produced from this calculation indicates the closest object. This label is then used along  
258 with the prediction for the performed action to infer which of the activity classes is being conducted  
259 within the frame. With the action, object and activity labels populated, the original frame is annotated  
260 as shown in Figure 6.



261

262 **Figure 6.** Annotated frame showing the person bending at the fridge

263 The annotated image shows the frame number in yellow, the predicted action in red, the nearest  
264 object in purple and the inferred activity in dark blue. In this frame the person is predicted to be  
265 bending at the fridge and so the *Using the Fridge* activity is inferred.

### 266 3. Experiment

267 For the experiment, each of the single component activities to be predicted were performed five  
268 times in a non-uniform order. This allowed us to adequately test the approach's capability to infer  
269 the correct activity, regardless of the order the activities were performed in. Both the TIS from the  
270 ceiling and from the side of the room were used for data capture. The thermal frames retrieved from  
271 both sensors during the performance of the activities were initially stored locally. This allowed the  
272 opportunity to create a ground truth for each of the frames prior to processing and performance  
273 evaluation.

274 This ground truth was created by processing each frame one at a time, along with the pairing  
 275 frame from the other TIS. The feature vectors for each frame in a pair were calculated, combined and  
 276 stored. Each feature vector was then manually labelled with the action being performed, object the  
 277 person was near, if any, and the activity that was being performed, if any. This provided a ground  
 278 truth state for each of the frames captured during the experiment. From each sensor 586 frames were  
 279 captured, making a total of 1172 thermal frames. There were, therefore, 586 feature vectors with a  
 280 size of 28. Table 3 presents how many frames were labelled with each of the actions, objects and  
 281 activities.

282 Once the ground truth was established, the accuracy of the system’s action, object and activity  
 283 recognition could be tested. For each frame from both TISs, the features were extracted and combined  
 284 to be passed through the trained random forest model. This produced a prediction for the action  
 285 being performed.

286 The proximity to objects within the room was also calculated to estimate whether the person was  
 287 within distance of the known position of an object that could be used. The value for the object was  
 288 determined as either, *Near Fridge*, *Near Coffee Cupboard*, or *Near Table*. The activity was inferred from  
 289 both the predicted action and object values, where it could have been one of four possible activities:  
 290 *Opening/Closing the Fridge*, *Using Fridge*, *Using Coffee Cupboard* or *Sitting At Table*.

291 When the predictions for each of the action, object and activity values were found, they were  
 292 each compared with the pre-established ground truth for that given frame to determine whether the  
 293 predictions were correct. Once each frame had been analysed, this allowed a total recognition  
 294 accuracy for each of the previously mentioned labels to be calculated.

295

**Table 3.** Number of frames containing each label

<b>Label</b>	<b>Number of Frames with Label</b>
<b>ArmsDown</b>	151
Rfwd	32
Lfwd	55
Rside	10
Lside	8
Bend	118
Sitting	212
Opening/Closing Fridge	27
Using Fridge	118
Using Coffee Cupboard	78
Sitting at Table	212
Near Fridge	148
Near Coffee Cupboard	78
Near Table	213

296

297 **4. Results**

298 In this Section we present the accuracy results achieved from training various machine learning  
 299 models. The prediction rates for the action performed, nearest object and inferred activities from the  
 300 conducted experiment are also broken down and evaluated.

301 *4.1 Models and Overall Results*

302 As stated previously, for each pair of frames from the two thermal sensors processed, a  
 303 prediction was made for the action, the object the person was near, and the single component activity  
 304 being performed. Where *S1* and *S2* are the frames from the ceiling and side sensor respectively, *F* is  
 305 the feature vector, *A* is the predicted action, *O* is the nearest object and *ADL* is the inferred activity,  
 306 the inference is displayed in Equation 1 and Equation 2.

$$S1 + S2 = F = A \tag{1}$$

$$A + O = ADL \tag{2}$$

307 For the prediction of the performed action, a machine learning algorithm was required. Of the  
 308 three models tested, the random forest model, in terms of training data accuracy, achieved the best  
 309 results. In Table 4, the accuracies for the action training data achieved by each model are presented.  
 310 These values are based on 10-fold cross-validation.

311 **Table 4.** Performance accuracies based on 10-fold cross-validation

Model	Action Accuracy (%)
Random Forest	97.10
Quadratic SVM	95.20
Complex Decision Tree	92.90

312 The models were then used in the experiment to analyse each frame and predict the action, detect  
 313 the object proximity and infer the activity. The results for the three models are shown in Table 5.

314 **Table 5.** Table showing results from each of the tested models

Model	Action (%)	Proximity (%)	Activity (%)
Random Forest	88.91	81.05	91.47
Quadratic SVM	68.40	81.05	74.20
Complex Decision Tree	86.68	81.05	91.29

315 The proximity accuracy does not change from model to model as it is not influenced by the  
 316 approach of the chosen machine learning algorithm. The threshold to determine what is and what is  
 317 not near is the only factor that plays a part in the proximity prediction. The activity prediction  
 318 accuracy, therefore, varies from model to model only because the action accuracy does. Even though  
 319 the activity accuracy achieved by the decision trees model is virtually identical to what is  
 320 accomplished by the random forest, it is the improvement in the action prediction accuracy that made  
 321 the random forest the best choice.

#### 322 4.2 Performed Action Results

323 During the experiment there were features extracted from the shape of the person's BLOB which  
 324 were used to predict the action the person was performing for that given frame. The results of these  
 325 predictions for each of the seven action classes are presented in Table 6. The *Rside* action appears to  
 326 be the worst performing action with a poor recognition rate. This is inverted with regards to the *Lside*  
 327 action as it was predicted correctly every time it was performed. This was almost achieved with the  
 328 *Bend* action as well as the *ArmsDown* action. This differentiation between *Bend* and *ArmsDown* was  
 329 made possible with the side sensor. This extra sensor data alleviated the burden on the ceiling sensor  
 330 to detect differences between the two actions, resulting in the two actions rarely being confused with  
 331 one another.

332 **Table 6.** Results for the predictions of the performed actions

Action	F-Score (%)	FPR (%)	FNR (%)	Precision (%)	Sensitivity (%)	Specificity (%)
ArmsDown	88.00	7.58	5.30	82.18	94.70	92.42
Bend	99.15	0.000	1.700	100.0	98.30	100.0
Lfwd	87.71	1.89	9.100	84.75	90.90	98.11
Lside	64.00	1.72	0.000	47.06	100.0	98.28
Rfwd	61.76	2.910	34.37	58.33	65.63	97.09
Rside	0.000	0.000	100.0	0.000	0.000	100.0
Sitting	92.42	0.2900	13.68	99.46	86.32	99.71

333 The low performance of *Rside* is again reiterated by the generated confusion matrix for the  
 334 actions in Table 7. In this table, the row shows the true action and each column shows the action that  
 335 was predicted. The rows show the actual number of instances for each action. The green box in the  
 336 rows demonstrate the number of times the action was correctly predicted (True Positive). The  
 337 columns show the number of times each action was predicted, either correctly or incorrectly. The  
 338 green box shows the number of correct predictions, while the red boxes show the times the action  
 339 was predicted, however, wrongly so (False Positive).

340 It can be hypothesised that the *Rside* performance was low due to the occlusion of the right arm  
 341 from the side sensor. Throughout the experiment the right and left arms were only ever extended out  
 342 to the side when the fridge or coffee cupboard were being opened. Due to the position of the side TIS,  
 343 the right arm was more likely to be occluded by the person's body, leaving the classification to only  
 344 the ceiling TIS. This could be addressed by capturing further frames of the *Rside* action being  
 345 performed to better train the ceiling sensor to classify this action on its own. The ceiling sensor may

346 have also struggled with the *Rside* action at the fridge as the fridge was quite low to the ground,  
 347 meaning the right arm was not required to extend to the side particularly far. The inference of the  
 348 activity did not suffer too much from this, as almost half of the misclassified *Rside* actions were  
 349 classified as the *Lside* action, which resulted in the same activity being inferred anyway.

350

**Table 7.** Confusion matrix created from the actions predictions

<i>True Class</i>	<i>Arms Down</i>	143			2	5		1
	<i>Bend</i>	2	116					
	<i>Lfwd</i>	2		50	3			
	<i>Lside</i>				8			
	<i>Rfwd</i>	7		4		21		
	<i>Rside</i>	3		2	4	1	0	
	<i>Sitting</i>	17		3		9		183
			<i>Arms Down</i>	<i>Bend</i>	<i>Lfwd</i>	<i>Lside</i>	<i>Rfwd</i>	<i>Rside</i>
		<i>Predicted Class</i>						

351 **4.3 Proximity Detection Results**

352 The person’s distance from each object’s proximity points was calculated to determine the object  
 353 the person was closest to, if they were within the specified threshold. The results for each object are  
 354 shown in Table 8.

355

**Table 8.** Results for the calculations of the proximity detection for any given frame

<b>Object</b>	<b>F-Score (%)</b>	<b>FPR (%)</b>	<b>FNR (%)</b>	<b>Precision (%)</b>	<b>Sensitivity (%)</b>	<b>Specificity (%)</b>
Fridge	87.57	11.38	0.000	77.89	100.0	88.62
Coffee Cupboard	85.71	5.210	3.850	77.30	96.15	94.79
Kitchen Table	90.06	15.21	0.000	81.90	100.0	84.79
None	41.94	0.000	73.47	100.0	26.53	100.0

356 The confusion matrix for the proximity detections that were produced from the experiment is  
 357 displayed in Table 9 and shows how the *None* label is main reason for lowering the accuracy value.  
 358 The person is frequently detected as being near the objects when actually, they are not near any of

359 them. This, however, does not affect the accuracy of the activity inference as the proximity detection  
 360 for the three objects is almost 100% accurate any time the person is actually near one of them.

361 **Table 9.** Confusion matrix created from the proximity detections

<i>True Class</i>	<i>Fridge</i>	148			
	<i>Coffee Cupboard</i>		75	3	
	<i>Table</i>			213	
	<i>None</i>	42	22	44	39
		<i>Fridge</i>	<i>Coffee Cupboard</i>	<i>Table</i>	<i>None</i>
	<i>Predicted Class</i>				

362 **4.4 Activity Inference Results**

363 From both the performed action and the nearest object to the person, the activity, if any, was  
 364 inferred. The results for the prediction of the performed activity within each frame are presented in  
 365 Table 10.

366 **Table 10.** Results for the predictions of the inferred activities for all frames captured during the experiment

<b>Activity</b>	<b>F-Score (%)</b>	<b>FPR (%)</b>	<b>FNR (%)</b>	<b>Precision (%)</b>	<b>Sensitivity (%)</b>	<b>Specificity (%)</b>
<b>Opening/Closing the Fridge</b>	80.00	0.5800	25.93	86.96	74.07	99.42
Using the Fridge	99.15	0.000	1.690	100.0	98.31	100.0
Using the Coffee Cupboard	94.59	0.00	0.2600	100.0	99.74	100.0
Sitting at the Table	92.42	0.2800	13.68	99.46	86.32	99.72
None	85.47	10.57	2.650	76.17	97.35	89.43

367 As stated, it was the results from the action classification and proximity detection from which  
 368 the activities were classified. The slightly lower proximity detection accuracy does not have any  
 369 significant detrimental effect on the activity accuracy. This was most likely because the  
 370 misclassifications of the nearest object were caused by the person walking past an object as opposed  
 371 to using one object, however, being predicted as near another. The low detection rate for the *Rside*  
 372 action also does not show any significant negative effects on the activity accuracy. The confusion  
 373 matrix for the activity predictions is presented in Table 11.

374

375

376

**Table 11.** The confusion matrix created from the activity predictions

True Class	Opening/Closing Fridge	20				7
	Using Fridge		116			2
	Using Coffee Cupboard			70		8
	Sitting At Table				183	29
	None	3			1	147
		Opening/Closing Fridge	Using Fridge	Using Coffee Cupboard	Sitting At Table	None
	Predicted Class					

377 **5. Discussion and Conclusions**

378 This aim of this paper was to propose an unobtrusive and accurate approach to single  
 379 component activity recognition. The study involved evaluating the use of two TISs for activity  
 380 recognition where it was found that the introduction of the second sensor benefited the accuracy of  
 381 using only TIS device types for activity recognition. We captured data for seven different actions to  
 382 train various machine learning models, where the random forest achieved the highest accuracy. The  
 383 positions of three objects within the kitchen were noted and action and object combinations were  
 384 determined to allow for the inference of single component activities. The trained model was tested  
 385 and evaluated to determine its ability to predict the actions and, as a result, the inferred activity.

386 The conducted experiment allowed for thermal frames to be captured to evaluate the trained  
 387 random forest model. A prediction for the performed action and the closest object were used in  
 388 conjunction with one another to infer if an activity was being performed in the frame. This was  
 389 completed for each of the frames, where the predictions were compared with the ground truth to  
 390 determine a recognition accuracy for each of the three labels. These experimental results were very  
 391 good with accuracies of 88.91%, 81.05% and 91.47% achieved for the action, proximity detection and  
 392 inferred activity, respectively. With the incorporation of the side sensor, actions such as *ArmsDown*  
 393 and *Bend* were easily distinguishable. The second sensor also helped avoid issues caused by image  
 394 noise, making the approach more robust. When too much noise caused difficulties in detecting the  
 395 person’s shape, making the frame unusable for extracting features, the frame could be disposed of  
 396 without concern as the second sensor’s frame could still be used on its own for feature extraction.

397 The *Rside* action prediction underperformed with each of its ten instances being misclassified as  
 398 another action. The implication of this low accuracy is, however, alleviated by the fact that almost  
 399 half of the misclassifications are for *Lside*, resulting in a correctly inferred activity anyway. This low  
 400 accuracy is also in the minority as the other targeted actions were predicted with high accuracy,  
 401 shown by the 100% and 99.46% precision values for *Bend* and *Sitting* respectively.

402 The results for the proximity detection was adequate, however, limited. The thresholds chosen  
 403 for the distances in the X and Y planes proved to be appropriate for attaining the best proximity



404 accuracy. This shows that there will be a need for refinement and further innovation in the proximity  
405 area of the work to subsequently improve upon the activity inference accuracy, potentially through  
406 the implementation of ultra-wideband (UWB) for 3D positioning of the kitchen objects. The activity  
407 inference yielded a high recognition accuracy supporting the case for the TIS device as an efficient  
408 and more than effective means for single component activity recognition within a smart environment.

409 This approach has, therefore, demonstrated that advantages of image processing techniques  
410 with visible spectrum images for smart home moderation can be retained, without breaching privacy,  
411 using only the TIS device. This is facilitated through its unobtrusive collection of data as no  
412 discernible characteristics of people are targeted, and through its automated nature as no wearable  
413 devices are required to monitor inhabitants. There is, however, potential for even further  
414 improvement and expansion of this method.

415 The need for future work to enhance the proposed system has been considered. While a more  
416 extensive set of training data could improve the accuracy of the *Rside* action, the issue may be one of  
417 occlusion. The prediction rate could then be improved by implementing an eighth action class,  
418 *Occluded*. This label would belong to frames where the ceiling sensor's feature data describes one  
419 action e.g. *Rside*, while the side sensor data describes another e.g. *ArmsDown*. In such scenarios, the  
420 frame and the feature data extracted from it would be disregarded for the inference of the performed  
421 activity.

422 The dataset used was imbalanced for some class labels, for both training and testing and  
423 although relatively high accuracies were achieved this imbalance will be addressed in future work.  
424 The imbalance was likely caused by the manner in which each action was captured. As a person is  
425 likely to perform each action randomly and for varying durations in a real-life scenario, the training  
426 data for a particular action was captured by performing that action in a similar vein. For example, if  
427 a five-minute time limit was used to capture some data for the *Lside* action, the person would perform  
428 this action in different parts of the room for different durations. The intention was that the training  
429 data would be made up of actions being performed in more realistic scenarios. This resulted in the  
430 data including frames of the person doing movements other than the targeted action such as walking  
431 and performing the *ArmsDown* action.

432 For the classes in the testing dataset, the experiment involved completing the activities five times  
433 each with no given time limit for the activity performance. This meant that the time spent on each  
434 activity was not necessarily equal, resulting in some actions being performed more than others. This  
435 inequality was also likely caused as some actions were not necessary for some activities, for example,  
436 *Sitting* was not required for using the fridge. A more balanced set of training data, however, may  
437 produce an even more accurate recognition rate. The approach to capturing training data in future  
438 work will therefore be stricter and more aimed toward a balanced class size rather than the recreation  
439 of a real-life scenario.

440 The system described in this study will be expanded upon in the future to not only recognise  
441 sub activities but the ADLs they make up. This will require an understanding of which sub-activities  
442 make up each targeted ADL and which actions signal their beginning and end. It will be vital to  
443 facilitate the tracking of the performed sub-activities over time to analyse the several activities that  
444 encompass the ADL performance, as opposed to the single frame analysis that is demonstrated here.  
445 With this, it will also be important to incorporate, for example, a Bayes statistical model to apply

446 probabilities to each of the activities potentially being performed. This will allow for evidence to be  
447 built over time to better determine the likelihood of an activity being performed. Different  
448 combinations of the list of extracted features may also be examined with the intention of efficiently  
449 improving the prediction rate of activities within a smart environment. Further sensor fusion  
450 approaches will be investigated, potentially involving other sensor types.

451 Maintaining privacy for inhabitants of smart environments remains an important factor in ADL  
452 analysis. Due to this, regardless of the future work that is conducted to improve upon the findings of  
453 this study, the preservation of the system's unobtrusive nature will remain a priority.

## 454 6. Patents

455 **Author Contributions:** Conceptualisation, Matthew Burns, Philip Morrow, Chris Nugent, Sally  
456 McClean; Data Curation, Matthew Burns; Methodology, Matthew Burns, Philip Morrow, Chris  
457 Nugent, Sally McClean; Software, Matthew Burns; Supervision, Philip Morrow, Chris Nugent, Sally  
458 McClean; Writing – original draft, Matthew Burns; Writing – review & editing, Matthew Burns,  
459 Philip Morrow, Chris Nugent, Sally McClean;

## 460 References

- 461 1. United Nations, Department of Economic and Social Affairs, P.D. *World Population Prospects:*  
462 *The 2017 Revision, Key Findings and Advance Tables. Working Paper No. ESA/P/WP/248; 2017;*
- 463 2. Farber, N.; Shinkle, D.; Lynott, J.; Fox-Grage, W.; Harrell, R. *Aging in place: A state survey of*  
464 *livability policies and practices; AARP Public Policy Institute, 2011;*
- 465 3. Rantz, M.J.; Skubic, M.; Miller, S.J.; Galambos, C.; Alexander, G.; Keller, J.; Popescu, M. Sensor  
466 Technology to support Aging in Place. *J. Am. Med. Dir. Assoc.* **2013**, *14*, 386–391.
- 467 4. Noury, N.; Poujaud, J.; Fleury, A.; Nocua, R.; Haddidi, T.; Rumeau, P. Smart Sweet Home...  
468 A Pervasive Environment for Sensing our Daily Activity? *Act. Recognit. Pervasive Intell.*  
469 *Environ.* **2011**, *4*, 187–208.
- 470 5. Alam, M.A.U.; Roy, N. GeSmart: A gestural activity recognition model for predicting  
471 behavioral health. In Proceedings of the 2014 International Conference on Smart Computing;  
472 IEEE, 2014; pp. 193–200.
- 473 6. Chen, B.; Fan, Z.; Cao, F. Activity Recognition Based on Streaming Sensor Data for Assisted  
474 Living in Smart Homes. *2015 Int. Conf. Intell. Environ.* **2015**, 124–127.
- 475 7. Nguyen, H.; Lebel, K.; Bogard, S.; Goubault, E.; Boissy, P.; Duval, C. Using Inertial Sensors to  
476 Automatically Detect and Segment Activities of Daily Living in People with Parkinson's  
477 Disease. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2018**, *26*, 197–204.
- 478 8. Gaglio, S.; Re, G. Lo; Morana, M. Human Activity Recognition Process Using 3-D Posture  
479 Data. *IEEE Trans. Human-Machine Syst.* **2015**, *45*, 586–597.
- 480 9. Barman, D.; Sharma, U.M. A Study On Human Activity Recognition From Video. In

- 481 Proceedings of the Proceedings of the 10th INDIACom; 2016 3rd International Conference on  
482 Computing for Sustainable Global Development, INDIACom 2016; M.N., H., Ed.; Institute of  
483 Electrical and Electronics Engineers Inc., 2016; pp. 2832–2835.
- 484 10. Liciotti, D.; Frontoni, E.; Zingaretti, P.; Bellotto, N.; Duckett, T. HMM-based Activity  
485 Recognition with a Ceiling RGB-D Camera. In Proceedings of the ICPRAM: Proceedings of  
486 the 6th International Conference on Pattern Recognition Applications and Methods; 2017; pp.  
487 567–574.
- 488 11. Baum, L.E. An Equality and Associated Maximization Technique in Statistical Estimation for  
489 Probabilistic Functions of Markov Processes. *Inequalities* 1972, 3, 1–8.
- 490 12. Pontes, B.; Cunha, M.; Pinho, R.; Fuks, H. Human-sensing: Low resolution thermal array  
491 sensor data classification of location-based postures. In Proceedings of the Lecture Notes in  
492 Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture  
493 Notes in Bioinformatics); Springer, Cham, 2017; Vol. 10291 LNCS, pp. 444–457.
- 494 13. Heimann Welcome to Heimann Sensor's Website Available online:  
495 [http://www.heimannsensor.com/products\\_imaging.php](http://www.heimannsensor.com/products_imaging.php) (accessed on May 28, 2018).
- 496 14. Nugent, C.D.; Mulvenna, M.D.; Hong, X.; Devlin, S. Experiences in the development of a  
497 Smart Lab. *Int. J. Biomed. Eng. Technol.* **2009**, 2, 319.
- 498 15. Rafferty, J.; Synnott, J.; Ennis, A.; Nugent, C.; McChesney, I.; Cleland, I. SensorCentral: A  
499 research oriented, device agnostic, sensor data platform. In Proceedings of the Lecture Notes  
500 in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture  
501 Notes in Bioinformatics); 2017; Vol. 10586 LNCS, pp. 97–108.
- 502 16. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man.*  
503 *Cybern.* **1979**, 9, 62–66.



© 2018 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license

506 (<http://creativecommons.org/licenses/by/4.0/>).