Journal of Sensor and Actuator Networks

MDPI

Article

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

Matthew Burns, Philip Morrow, Chris Nugent and Sally McClean

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

Received: date; Accepted: date; Published: date

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

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

---

1. Introduction

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

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

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

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

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

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

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

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

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

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

2. Materials and Methods

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

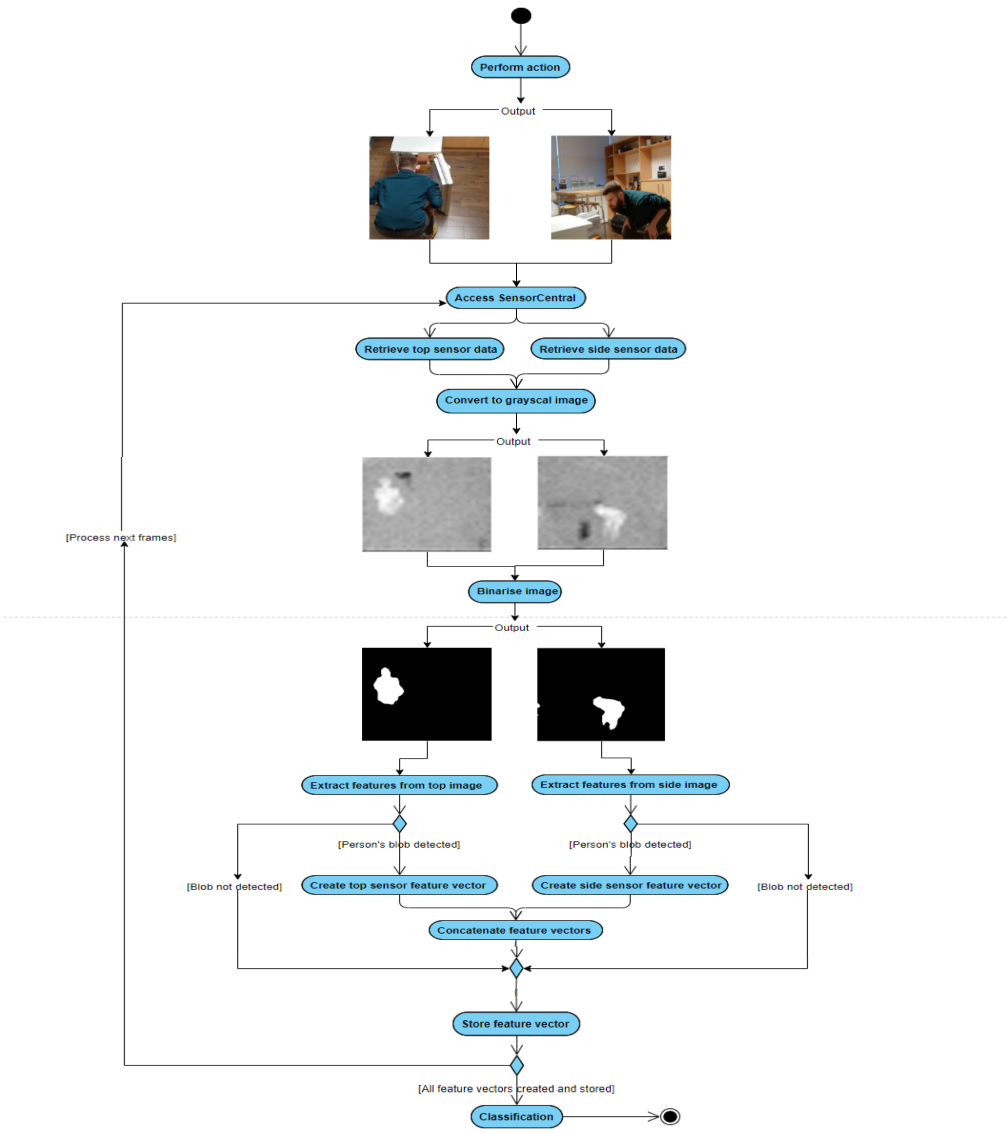


Figure 1. Overview of the initial stages of the method.

The fundamental functionality of this single component activity recognition approach is to retrieve thermal frames from two sensors of the same type and extract and fuse relevant features to predict the single component activity being performed within each frame. Upon determination of the action being performed within the frame, the object that the person is nearest to is calculated. This 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
```

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

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

The first step in the process is to retrieve the thermal frames from *SensorCentral*, which acts as the middleware for the devices and the developed system. The raw data captured by the TIS is packaged in JSON format and consists of the frame data, timestamp and the sensor ID. The JSON formatted frame data from both TIS devices is retrieved and used to fill a 32x32 matrix. For convenience, the image is then resized to a 256x256 image. The TISs are, however, 32x31 sensors and so this 32nd row is simply a black line of pixels which when the image is resized to 256x256, makes up the bottom seven rows. These rows are removed, resulting in a 256x249 image. Once the frames from both sensors are established, they are binarised using Otsu's automatic threshold method [16]. This allows the person's shape to be analysed and features extracted to train the chosen machine learning model. Frames from both TISs are captured at the same time and upon retrieval of a pair of these frames, their timestamps are compared to ensure the frames were captured at the same instant and not seconds or more apart.

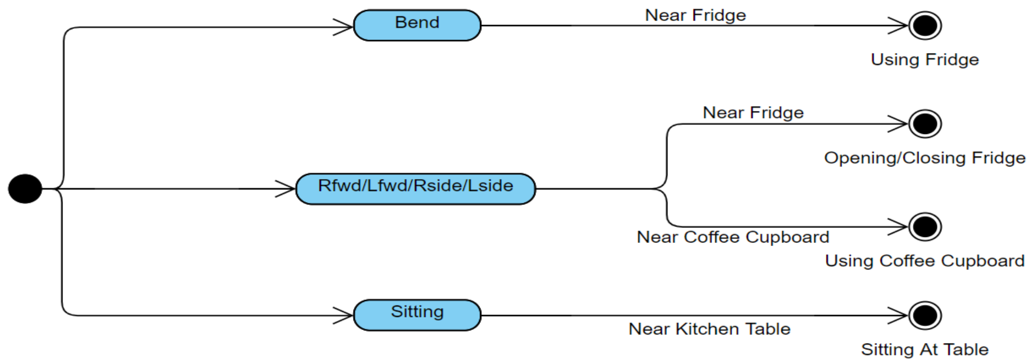


Figure 3. Overview of the activity inference process.

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

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

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

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

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

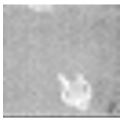
Feature	Description
Eccentricity	The ratio of the distance between the foci of the shape's ellipse and its major axis length
Major and minor axis ratio (Pixels)	Ratio between the length of the major axis of the ellipse and the length of the minor axis of the ellipse
Standard Deviation	Standard Deviation 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 *x*-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

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.

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.

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.

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		

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.

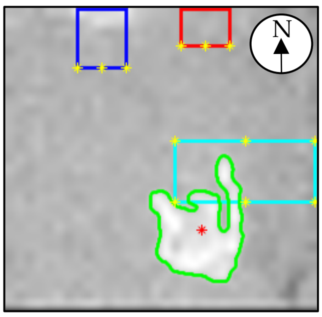


Figure 4. A person sitting at the kitchen table, as seen by the ceiling TIS

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

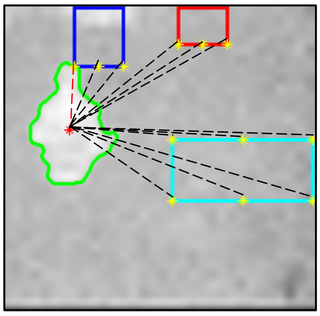


Figure 5. Depiction of the distance measurement between the person's centroid and each object's proximity points

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

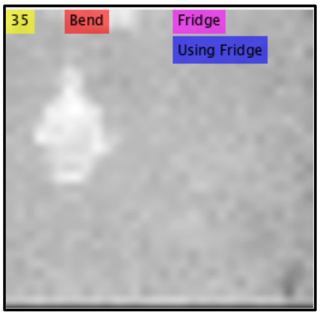


Figure 6. Annotated frame showing the person bending at the fridge

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

3. Experiment

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

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

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

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

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

Table 3. Number of frames containing each label

Label	**Number of Frames with Label**
ArmsDown	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

4. Results

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

4.1 Models and Overall Results

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

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

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

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

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

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

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

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

4.2 Performed Action Results

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

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

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

It can be hypothesised that the *Rside* performance was low due to the occlusion of the right arm from the side sensor. Throughout the experiment the right and left arms were only ever extended out to the side when the fridge or coffee cupboard were being opened. Due to the position of the side TIS, the right arm was more likely to be occluded by the person's body, leaving the classification to only the ceiling TIS. This could be addressed by capturing further frames of the *Rside* action being performed to better train the ceiling sensor to classify this action on its own. The ceiling sensor may have also struggled with the *Rside* action at the fridge as the fridge was quite low to the ground, meaning the right arm was not required to extend to the side particularly far. The inference of the activity did not suffer too much from this, as almost half of the misclassified *Rside* actions were classified as the *Lside* action, which resulted in the same activity being inferred anyway.

Table 7. Confusion matrix created from the actions predictions

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

4.3 Proximity Detection Results

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

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

Object	**F-Score (%)**	**FPR (%)**	**FNR (%)**	**Precision (%)**	**Sensitivity (%)**	**Specificity (%)**
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

The confusion matrix for the proximity detections that were produced from the experiment is displayed in Table 9 and shows how the *None* label is main reason for lowering the accuracy value. The person is frequently detected as being near the objects when actually, they are not near any of them. This, however, does not affect the accuracy of the activity inference as the proximity detection for the three objects is almost 100% accurate any time the person is actually near one of them.

Table 9. Confusion matrix created from the proximity detections

True Class \ Predicted Class	*Fridge*	*Coffee Cupboard*	*Table*	*None*
Fridge	148			
Coffee Cupboard		75	3	
Table			213	
None	42	22	44	39

4.4 Activity Inference Results

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

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

Activity	F-Score (%)	FPR (%)	FNR (%)	Precision (%)	Sensitivity (%)	Specificity (%)
Opening/Closing the Fridge	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

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

Table 11. The confusion matrix created from the activity predictions

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

5. Discussion and Conclusions

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

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

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

The results for the proximity detection was adequate, however, limited. The thresholds chosen for the distances in the X and Y planes proved to be appropriate for attaining the best proximity accuracy. This shows that there will be a need for refinement and further innovation in the proximity area of the work to subsequently improve upon the activity inference accuracy, potentially through the implementation of ultra-wideband (UWB) for 3D positioning of the kitchen objects. The activity inference yielded a high recognition accuracy supporting the case for the TIS device as an efficient and more than effective means for single component activity recognition within a smart environment.

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

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

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

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

The system described in this study will be expanded upon in the future to not only recognise sub activities but the ADLs they make up. This will require an understanding of which sub-activities make up each targeted ADL and which actions signal their beginning and end. It will be vital to facilitate the tracking of the performed sub-activities over time to analyse the several activities that encompass the ADL performance, as opposed to the single frame analysis that is demonstrated here. With this, it will also be important to incorporate, for example, a Bayes statistical model to apply probabilities to each of the activities potentially being performed. This will allow for evidence to be built over time to better determine the likelihood of an activity being performed. Different combinations of the list of extracted features may also be examined with the intention of efficiently improving the prediction rate of activities within a smart environment. Further sensor fusion approaches will be investigated, potentially involving other sensor types.

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

6. Patents

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

References

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

Proceedings of the Proceedings of the 10th INDIACom; 2016 3rd International Conference on Computing for Sustainable Global Development, INDIACom 2016; M.N., H., Ed.; Institute of Electrical and Electronics Engineers Inc., 2016; pp. 2832–2835.

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