2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Learned Sensor Fusion For Robust Human Activity Recognition in Challenging Environments

Max Conway[1], Brian Reily[2] and Christopher Reardon[1]

***Abstract*— Human activity recognition is a vital area of robotics with significant real-world applications, from enhancing security and surveillance to improving healthcare and human-robot interaction. A critical challenge lies in bridging the gap between research models, which often assume ideal conditions, and the complexities of real-world environments. In practice, conditions can be far from perfect, including scenarios with poor lighting, adverse weather, or blurred views. In this paper, we present an innovative approach for robust activity recognition through learned sensor fusion, in which our recognition framework identifies a latent weighted combination of input modalities, enabling classifiers to capitalize on advantages provided by various sensors. In support of our work, we have released a dataset of human activities across multiple modalities with environmental degradation factors such as darkness, fog, and thermal blur. Our proposed approach identifies a weighted combination of modality representations derived from existing architectures. We show that our approach is able to achieve 24% higher classification performance than existing single-modality approaches. Our approach also attains comparable performance to modality fusion approaches in significantly reduced classification time. In real-world robotics applications, particularly those occurring in dangerous, degraded environments, this speed is critical.**

I. Introduction

Human activity recognition (HAR) has emerged as a pivotal field at the intersection of robotics, computer vision, machine learning, and sensor technology, with implications across diverse domains [1]. Activity recognition enables robotic response to human activities based on data from sensors. HAR is a critical component in the future of human-computer interaction [2]–[4]. In healthcare, HAR can enhance the quality of life for the elderly and individuals with medical conditions through continuous activity monitoring, enabling early detection of health issues. In the realm of security and surveillance, HAR aids in identifying suspicious behavior, enabling more efficient response to threats. In disaster response, where robots are becoming crucial for use in dangerous environments, HAR enables non-verbal communication between humans their robotic teammates.

A major limitation of existing state-of-the-art HAR approaches is their reliance on the assumption that data is pristine, that is, often recorded in a well lit room with no obstructions [5]. These assumptions leave a challenging gap between research models and practical applications. Real-world environments often introduce challenges such as varying lighting conditions, occlusions, and environmental degradation, which can significantly impact the performance of a HAR system. In order to enable effective human-robot interaction in scenarios such as disaster response, robust human activity recognition that can both integrate multiple sensing modalities and weight them by their relevance is critical.

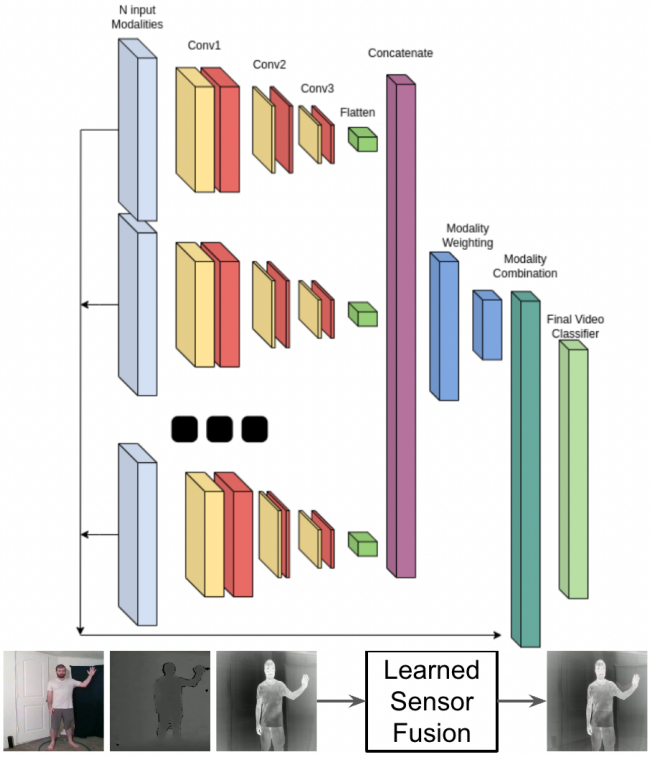


Fig. 1. Our proposed approach for robust human activity recognition enables the learned sensor fusion of multiple modalities, such as RGB, depth, and thermal. We propose a novel network architecture in which multiple modalities are selected and fused in a unified process with classification.

Existing approaches for HAR can be categorized into two general areas. First, skeleton based approaches use a package such as OpenPose [6] to extract skeletal features from an input stream. These skeleton features can then be used for classification. While accurate skeletal pose data can provide effective activity recognition performance [7]–[10], skeletal pose data packages are commonly geared towards only RGB or depth data, and need to be retrained to perform properly with other sensor modalities [5]. The second area focuses on activity recognition from deep approaches [11]–[15], which

[1]Department of Computer Science, University of Denver. {max.conway, christopher.reardon}@du.edu.
[2]DEVCOM U.S. Army Research Laboratory. brian.j.reily.civ@army.mil.

typically requires large amounts of data to be effective. However, these deep approaches have not been effectively demonstrated on data from challenging environments, where conditions can drastically degrade sensor performance, or with smaller training datasets that are typically available in robotics applications.

In this paper, we propose a more efficient alternative to traditional deep approaches, in which we are able to learn effective sensor fusion via modality selection unified with classification. We first create a novel dataset of human activities in a variety of challenging environmental conditions. In this dataset, activities are not only recorded in ideal conditions in full-color RGB, depth, and thermal modalities; but we also degrade conditions by introducing fog, removing lighting, or adding thermal blur. We then design a unified network structure to learn an intermediate weighting of modalities as we perform classification with an existing state-of-the-art classifier. An overview of this process can be seen in Figure 1. Our proposed approach enables activity recognition without *a priori* identification of relevant modalities, enabling real-time operation in real-world degraded environments with dynamic and seamless sensor fusion based on the input data. In real-world robotics applications, particularly those that could occur in degraded environments such as search and rescue, efficient computation time can be critical for the safety of the robot and the humans it is interacting with.

We introduce two contributions in this paper:

1) First, we introduce a case-study dataset of human activity recognition in challenging environments. This dataset consists of eight different actions performed in a variety of adverse conditions (e.g., darkness, fog, thermal blur). This dataset is available publicly and is believed to be among the first of its kind for human activity recognition in challenging environmental conditions.
2) Second, we propose a novel approach to human activity recognition in challenging environments. Our approach integrates dynamic modality selection for sensor fusion based on the input data directly into the classification process, improving classification versus single modalities or equal-weighted modalities while significantly improving processing time versus ensemble methods.

II. Related Work

In the area of Human Activity Recognition (HAR), research has predominantly gravitated towards two key modalities: RGB and RGBD. The former leverages the visual richness of standard RGB images and videos to infer human activities, while the latter extends this by incorporating depth information, often extracted from sensors such as the Microsoft Kinect or Intel Realsense. Furthermore, HAR classification methods can be categorized into two main groupings: skeleton-based methods that rely on precise skeletal representations extracted from images or depth data [7], [10] and deep learning-based methods that directly process visual data to infer activities [11], [16], [17]. While approaches in both categories have become highly accurate under ideal conditions, challenges still exist across HAR research, such as occlusions [9], subject segmentation [6], similarities between actions [3], and illumination conditions [7], [18].

A. Activity Recognition Methods

Skeletal data extraction is a common preprocessing step for human activity recognition. Pose estimation approaches [5], [6], [9] operate on sensor data, typically RGB or RGBD, and return a representation of the human form as joint positions [8]. Fundamental challenges of skeletal data extraction include the environmental conditions of the scene such as clutter, occlusions and lack of illumination [5]. One solution to make skeletal data extraction possible for a variety of lighting conditions is to incorporate other sensing modalities, such as thermal camera imagery [5], [19], [20] or wearable technology measurements [21]. Many approaches exist for classifying human actions from the resulting skeleton representations, such as multiclass SVM [7], regularized convex optimization [22], [23], deep recurrent neural networks (RNN) for sequences actions [18], fusing observations of multiple individuals [24], or action graphs to characterize postures [10]. Skeletal representation extraction can be computationally expensive, and so is better suited for individual modalities when real-time performance on a robot is needed.

Deep approaches have shown incredible accuracy in general video classification problems. Many different models exists to classify videos such as Movinet [11], Efficientnet [12], and convolutional neural networks [14]. Many of these models are available pre-trained on datasets such as Kinetics-600 [17], which makes them appropriate for transfer learning to human activity recognition applications. However, as with approaches that rely on the extraction of skeletal data, deep approaches have largely been used solely on RGB and RGBD data, and can become quite computationally expensive when additional modalities increase the size of networks or the number of separate networks. Even state-of-the-art approaches such as UMDR [15], which is able to combine information from both RGB and depth sensors, are not lightweight enough to deploy on mobile robots.

B. Activity Recognition Datasets

In support of this variety of HAR research, a wide variety of human activity datasets have been introduced. The most popular RGBD datasets for evaluating human activity recognition performance are the MSR Action 3D dataset [10], the Cornell Activity Dataset [2], and the NTU RGBD dataset [3]. These datasets all feature labeled RGBD recordings of different actors performing different actions. There are also popular datasets for 2D human activity recognition such as Kinetics [17] and Charades [16], which both provide RGB videos. While these popular datasets have proved useful in pushing forward HAR research, there is a lack of existing datasets that include actions recorded in additional sensor

modalities (e.g., thermal) or datasets that are recorded in challenging environmental conditions. Motivated by this, we contribute a dataset where actions are recorded in RGB, thermal, and depth modalities, with a variety of environmental and illumination challenges introduced.

C. Thermal Cameras and Sensor Fusion

Thermal cameras have shown great promise for robust perception. Thermal cameras have already been shown to perform better for perception of surroundings than RGB or LiDAR in disaster environments consisting of dense fog and low light [25]. Thermal sensors are also a promising solution to some privacy concerns [5], which is relevant when dealing with datasets of human actions. In many works, thermal information has been fused with other sensing modalities to increase performance. For example, the fusion between thermal and depth cameras has shown potential use for patient monitoring [19]. Fusion between thermal, RGB, and depth has been applied even more extensively, including works addressing object detection [26], semantic segmentation [27], pedestrian detection [28], object tracking [29], person re-identification [30], SLAM [31], building mapping [32], scene reconstruction [33], and facial recognition [34] [1], [20].

The fusion of multiple sensors is shown to improve the model performance for robust HAR. A frequent combination is the fusion of RGB and depth information [2], [3], [10]. Some approaches to human activity recognition use no visual modalities and instead rely on multiple wearable sensors [35], [36], while other studies use a mixture of visual modalities and worn sensors [4]. As no sensor is perfect and all will fail under the certain conditions, a method to dynamically change priority between sensors depending on the conditions has a potentially impactful opportunity to enable more robust HAR when degraded conditions occur.

III. Proposed Approach

A. Motivation

Modern robotics enables input from multiple different sensors which can provide valuable information for perception algorithms in a variety of environments. For example, RGB cameras, a widely-available and affordable sensor option, can be used in many non-degraded environments. However, as lighting conditions degrade, perception approaches reliant on this sensor type suffer performance decreases or fail altogether. Similarly, when thermal noise is present such as from engines or fires, or when materials such as insulated glass impede sensors, thermal data may become unreliable. When smoke or fog is present in the environment, occlusion and specular reflection render RGB and active depth sensors (e.g., LiDAR) ineffective. In these situations, effective prioritization of sensing modality information can overcome adverse environmental conditions. While it may be possible to manually determine the best sensor combination for individual situations, in real-world scenarios with a varying mixture of degraded conditions, intelligent modality selection through learned sensor fusion is essential to effective performance of any robotics application.

B. Problem Definition

In our proposed approach, we consider input from M modalities. We define an input modality stream as a 3-channel tensor based on the input video width, height, and number of frames. We consider all input tensors to be 3-channel, in order to accommodate the standard structure of the RGB modality. For modalities that are not intrinsically 3-channel, such as depth, we extend them by duplication of channels (i.e., a single depth channel data is copied into all 3 channels). Specifically, the i-th input to our proposed approach is a tensor from the m-th modality defined as $\mathbf{T}_i^m \in \mathbb{R}^{N \times W \times H \times 3}$, where N is the number of frames and W and H are the width and height of the frame in pixels.

We define a multi-layer network in which the output from a given video input is an activity classification. Given an input video $\mathbf{T}_i^m$, we aim to identify an output $\mathbf{y}_i \in \mathbb{R}^C$, a vector containing probabilities that the i-th input instance belongs to one of C classes. Then the predicted output class y_i^c can be identified as the highest probability in $\mathbf{y}_i$. However, our main contribution is that we propose a novel sensor fusion-based approach to learn a set of latent modality selection weights $\mathbf{w} \in \mathbb{R}^M$.

C. Proposed Network Structure

In order to attain not only the necessary activity recognition performance of compared approaches but also the introduced modality selection, we propose a novel network structure. This network architecture is visualized in Figure 1, with further detail shown in Figure 2.

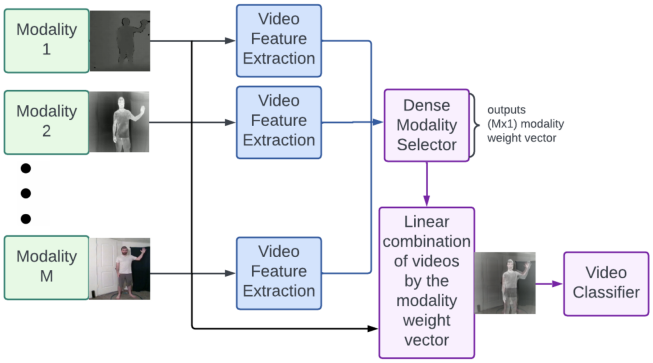


Fig. 2. Flow chart of our proposed approach. Video streams from multiple modalities are used as input to individual video feature extraction modules consisting of convolutional and dense layers. The extracted features feed into a dense modality selector, which identifies a weighting for sensor fusion. These weights are used to generate a linear combination of the original inputs, and a Movinet [11] classifier is used for the final activity recognition.

Our approach takes as input video streams from multiple modalities. Each video stream is processed by a proposed feature extraction module. The feature extraction consists of a series of convolutional and pooling layers to decrease the input size. These finally feed into individual flattening layers, which serve to compress the multidimensional data (across frames, width, height, and channels) into a one-dimensional representation. This representation is concatenated and processed by two dense layers which serve as the modality

selection module. This module takes all processed modalities as input and outputs $\mathbf{w} \in \mathbb{R}^M$, a one-dimensional weight vector where w_m represents the importance of the m-th modality. The vector $\mathbf{w}$ is obtained by applying the softmax function to ensure that the weights are non-negative and sum to one.

The weight vector $\mathbf{w}$ is then used to perform sensor fusion by generating a linear combination of the original input video streams:

$$\mathbf{T}_i^* = \sum_m^M w_m \mathbf{T}_i^m \tag{1}$$

This fused representation $\mathbf{T}_i^*$ can be used with any activity recognition classifier, which in our implementation is the Movinet classifier [11]. This classifier outputs the final class probability vector $\mathbf{y}_i$, which is used to identify the output class y_i^c.

By proposing an activity recognition approach that integrates modality selection with the classification framework, we are able to learn sensor fusion weights implicitly from the final activity labels. That is, instead of explicitly providing modality selection weights to the selection module (e.g., enforcing that the depth modality is not used in foggy conditions), we can efficiently learn these weights through this unified approach during the optimization process. Not only does this enable learned sensor fusion, but also means that only a single classifier is needed for multiple modalities, creating a more efficient recognition approach. These learned weights also provide explanatory power to our proposed approach, indicating which modalities the model is relying on as it learns and performs classification.

IV. Dataset

In support of the evaluation of our proposed approach, we introduce a case-study dataset of human activities in challenging environments[1]. While many human activity recognition datasets exist, as discussed in Section II, there is currently a lack of datasets that address thermal cameras, challenging environmental conditions, or both.

Our introduced dataset is recorded under four conditions: *Ideal*, with no introduced environmental challenges; *Dark*, in which no lights are turned on; *Fog*, where a fog machine was used to fill the room with fog; and *Thermal Blur*, where clear tape was placed over the thermal camera to simulate blurring that can occur in a real world deployment. RGB images of these conditions can be seen in Figures 3(a), 3(e), 3(i), and 3(m).

Our dataset consists of eight different actions or activities: *Clap*, *Draw Circle*, *Draw X*, *Point*, *Stand*, *Swipe Left*, *Swipe Right*, and *Wave*. These actions were chosen as they are representative of non-verbal commands that occur in disaster response scenarios, a key application of our proposed activity recognition for challenging environments. For example, a human first responder may *Point* to direct a robot to examine an area, or *Draw Circle* to gather their teammates. The actions are also inspired by those in the UTD-MHAD [4] dataset to remain consistent with existing practices.

[1] https://bit.ly/robust_har

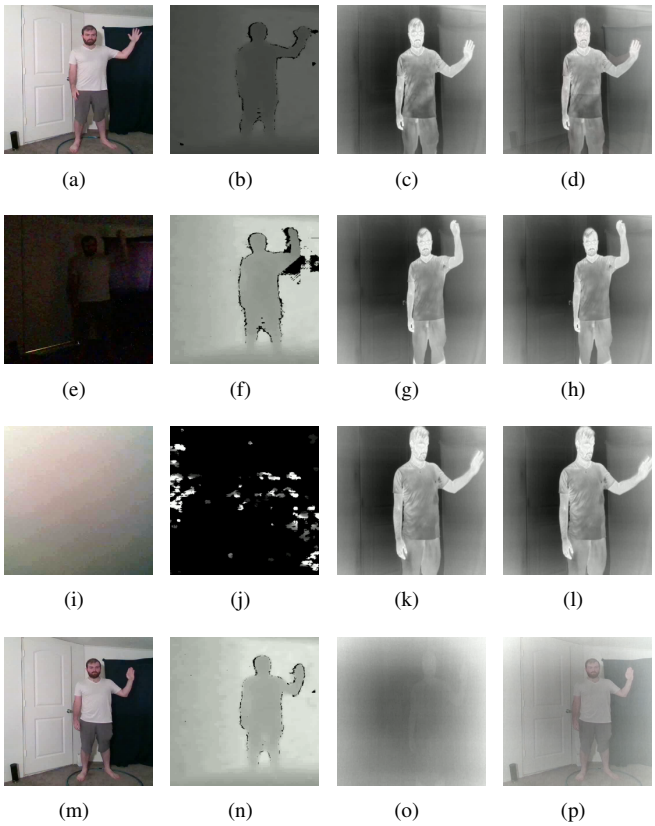


Fig. 3. Example frames from the introduced dataset of the *Wave* activity. The columns from left to right show the RGB modality, depth, thermal, and a fused frame utilizing weights learned by our approach. Figures 3(a)-3(c) are recorded in the *Ideal* condition, Figures 3(e)-3(g) are in the *Dark* condition, Figures 3(i)-3(k) are in the *Fog* condition, and Figures 3(m)-3(o) are in the *Thermal Blur* condition.

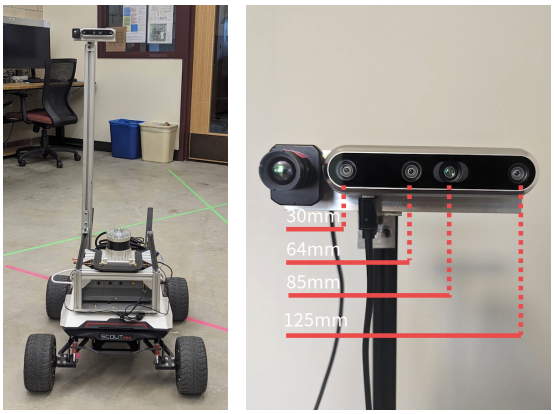
The dataset was recorded utilizing the AgileX Scout Mini and camera setup shown in Figure 4. An Intel Realsense camera provided the RGB and depth modalities, while a Flir Boson 640 was used to record the thermal modality. A calculated physical transform, seen in Figure 4(b), was used to align the images. Each action was recorded five times by two individuals in each of the four conditions, for a total of 320 video instances.

V. Experimental Results

A. Experimental Setup

To evaluate our approach, we compared to several variations based on the Movinet classifier [11] utilized as the last step in our proposed approach. Each Movinet classifier was initialized with weights from pre-training on the Kinetics-600 dataset [17]. Specifically, we compare our performance against:

1) *Individual Modalities*: As a baseline comparison, we trained an individual Movinet classifier for each recorded modality (*Ind-RGB*, *Ind-Depth*, and *Ind-Thermal*).
2) *Equal Weighted Modalities*: This comparison utilizes all available modalities but performs sensor fusion



(a) Robot Setup (b) Camera Setup

Fig. 4. The experimental setup used to record the introduced dataset. A AgileX Scout Mini was used as a mobile platform with a Flir Boson 640 and Intel Realsense mounted approximately $1m$ high.

TABLE I
OVERALL RESULTS

Method	Accuracy	Evaluation Time (Per Instance)
Ind-RGB	71.09%	0.032s
Ind-Depth	68.75%	0.031s
Ind-Thermal	78.90%	0.032s
UMDR RGB	68.75%	**0.021s**
UMDR Depth	80.46%	**0.021s**
UMDR Fusion	85.15%	0.039s
Equal Weighted	73.43%	0.034s
Modality Ensemble	**92.97%**	0.058s
Our Proposed Approach	**92.97%**	0.037s

through an equally weighted linear combination of inputs (i.e., the modality selection portion of our approach is not done, and $\mathbf{w}$ is defined such that each element $w_m = \frac{1}{M}$).

3) *Modality Ensemble*: This comparison method utilizes a Movinet classifier on each modality, then concatenates the resulting classification confidences and uses this to output an overall classification.
4) *UMDR*: Finally, we compare against the state-of-the-art Unified Multimodal De- and Re-Coupling (UMDR) framework [15]. This approach is able to consider input from both RGB and depth sensors. For this comparison method, we report its performance on only RGB, only depth, and when incorporating both (*UMDR Fusion*).

All approaches considered video instances where $N = 15$ (that is, each input video is 45 frames and every third frame is used).

For each comparison method and our proposed approach, we report the overall accuracy, as well as the time taken to classify a new input instance. We also report the accuracy of each method within each challenging environmental condition. Finally, for our proposed approach we present confusion matrices to understand difficulties in the dataset and report learned sensor fusion weights across the varying conditions. Training and evaluation for our proposed approach and each compared method was performed on a machine equipped with an Intel i7-12700H CPU and an Nvidia GeForce RTX 3060 Laptop GPU.

B. Training

With 10 total instances per action and condition, we utilized five instances for training (instances 1, 2, 3 from subject 1 and instances 6, 7 from subject 2). We used one instance for validation (instance 4 from subject 1). We used four instances for test evaluation (instance 5 from subject 1 and instances 8, 9, 10 from subject 2). In order to prevent overfitting on this compact case-study dataset, we utilized a variety of data augmentation methods to ensure that each method is able to learn from a large variety of training inputs. Specifically, we implemented data augmentation by adjusting *rotation*, *brightness*, and *shearing*:

1) *Rotation*. Videos are rotated by an angle of $\theta = \{-45°, \ldots, 45°\}$ by multiplying each frame by the rotation matrix $\mathbf{R}$:

$$\mathbf{R} = \begin{bmatrix} \alpha & \beta & (1-\alpha)x_c - \beta y_c \\ -\beta & \alpha & \beta x_c + (1-\alpha)y_c \end{bmatrix}$$
$$\alpha = \cos(\theta), \beta = \sin(\theta)$$

where (x_c, y_c) denotes the center of the frame.
2) *Brightness Adjustment*. Each frame of a video is brightness adjusted by a factor of $F_b = \{0.5, \ldots, 2\}$ by multiplying each frame by this factor:

$$\bar{\mathbf{T}_i^m} = \mathbf{T}_i^m * F_b$$

3) *Shearing*. Finally, each frame is sheared by a factor of $F_s = \{-0.5, \ldots, 0.5\}$. The shear matrix $\mathbf{S}$ is defined as:

$$\mathbf{S} = \begin{bmatrix} 1 & F_s & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Each pixel (x, y) is adjusted using this matrix:

$$(\bar{x}, y) = (\mathbf{S}_{11}x + \mathbf{S}_{12}y + \mathbf{S}_{13},\ \mathbf{S}_{21}x + \mathbf{S}_{22}y + \mathbf{S}_{23})$$

Training was performed for 32 epochs using a standard ADAM optimizer with a learning rate of $1e-4$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.

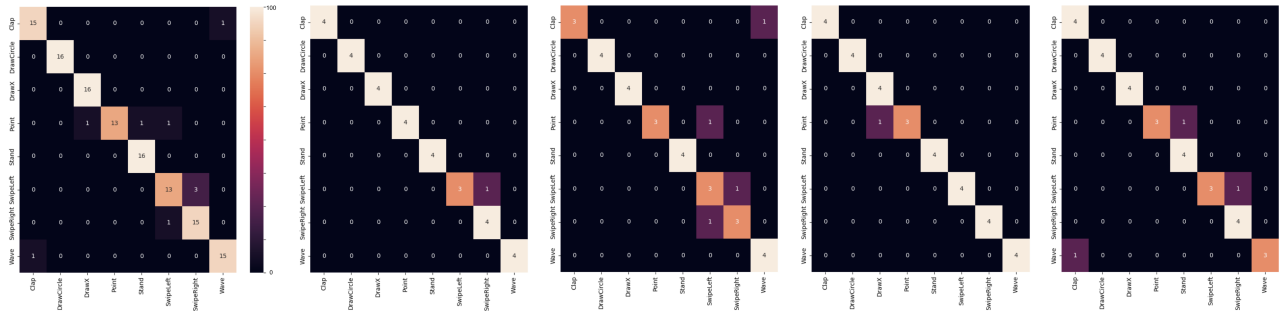
C. Evaluation Results

We present performance results in Table I across all of the challenging environmental conditions in our dataset. First, we can see that classifiers based on only a single sensing modality - regardless of whether it is RGB, depth, or thermal - attain subpar performance. In the challenging environmental conditions we introduce, each modality is particularly limited by a specific condition, causing it to struggle overall. We also see that the naive *Equal Weighted* combination of modalities performs similarly, with the equal linear combination of modalities causing irrelevant information to be included in the combined representation. The UMDR classifiers, though able to demonstrate state-of-the-art performance on external datasets and operate in the fastest time, also struggle on our challenging dataset when trained with only a single sensing modality. UMDR is able to learn from both RGB and depth data, and we do see that combining these modalities in *UMDR Fusion* does increase its performance. However, we

TABLE II

RESULTS BY ENVIRONMENTAL CONDITION

Condition	Ind-RGB	Ind-Depth	Ind-Thermal	UMDR Fusion	Equal Weighted	Modality Ensemble	Our Proposed Approach
Ideal	96.88%	81.25%	96.88%	100.00%	84.38%	96.88%	96.88%
Fog	21.88%	34.37%	87.50%	59.37%	46.88%	87.50%	87.50%
Dark	81.25%	81.25%	90.63%	84.37%	84.38%	100.00%	96.88%
Thermal Blur	84.38%	78.13%	40.63%	96.87%	78.13%	87.50%	90.63%
Non-Ideal	62.50%	64.58%	72.92%	80.20%	69.80%	91.67%	91.67%



(a) Overall (b) Ideal (c) Fog (d) Dark (e) Thermal Blur

Fig. 5. Confusion matrices for our proposed approach. Figure 5(a) shows overall results, while Figures 5(b)-5(e) show results for individual environmental conditions.

TABLE III

LEARNED SENSOR FUSION WEIGHTS

Condition	RGB	Depth	Thermal
Ideal	52.4%	2.2%	45.4%
Fog	15.2%	0.5%	84.3%
Dark	22.1%	1.8%	76.1%
Thermal Blur	55.1%	2.9%	42.0%
Overall	36.2%	1.9%	62.0%

see the best results in the approaches able to incorporate all three available modalities. The *Modality Ensemble*, by running three full classifiers plus additional concatenation layers, takes the longest time to evaluate a new video instance. In contrast, our proposed approach is able to achieve the same performance but in a similar amount of time as single modality classifiers. Specifically, our approach completes both training and individual instance classifications in 63% of the time that *Modality Ensemble* requires. In time sensitive applications like search and rescue situations, where robots must respond quickly to human teammates, this could prove crucial to mission success.

In Table II, we report the accuracy performance of each compared approach in each of the challenging environmental conditions introduced in our contributed dataset. Across all compared approaches, performance in *Ideal* conditions is high ($> 80\%$, with all but two approaches $> 90\%$). However, when the environmental conditions become challenging, performance begins to degrade. Specifically, we see that each *Individual Modality* method struggles with a particular environmental condition. *Fog* obscures both RGB and depth sensors, resulting in the two worst performances for the individual approaches relying on those. The *Equal Weighted* approach, unable to adapt and rely on more informative sensors, also struggles significantly. Even the *UMDR Fusion* approach, which attains state-of-the-art performance on external datasets and perfect accuracy in our *Ideal* conditions, performs very poorly when dealing with *Fog*.

The *Dark* condition is not nearly as challenging, with even *Ind-RGB* able to attain decent performance as we do not reduce the amount of light completely for this condition. Similarly, in the *Thermal Blur* condition only the *Ind-Thermal* approach struggles significantly.

When looking at challenging conditions individually and as a whole (with *Non-Ideal* encompassing results across every condition except *Ideal*), the two methods that attain the best results are the ones able to learn from multiple sensing modalities while also learning to balance among them: *Modality Ensemble* and our proposed approach. While the ensemble is able to perform as well as our method, we again note from Table I that our approach is able to do this in just over half the training and evaluation time.

In Table III, we report the latent sensor weights learned by our approach. The reported weights are the average of learned sensor weights after training. It is straightforward to see the connection to the performance of classification on individual modalities, which provides explainable power to our proposed model. Depth has the worst performance, and our approach has identified this by weighting the depth modality less than 3% in all conditions. RBG and thermal are also weighted in order of their overall performance. We can see that thermal is weighted over 75% in both *Fog* and *Dark*, the two conditions most challenging for RGB cameras. Similarly, the RGB weight is highest in the *Thermal Blur* condition, the only environmental challenge that significantly degrades the performance of the thermal modality.

Finally, Figure 5 presents confusion matrices for our approach overall and in each of the challenging environmental conditions. Each confusion matrix shows the true label on the left and the predicted label on the bottom. We see that the most difficult actions to differentiate are those of *Point* and *Swipe Left*, caused by the similar arm movement.

Multiple actions, specifically *Stand* and the *Draw* actions, are classified correctly in all conditions.

VI. Conclusion

Human activity recognition is crucial for effective human-robot interaction. However, despite wide-ranging research, there still exists a significant challenge in bridging the gap between training on ideal conditions from a laboratory setting and the complex and degraded environments that exist in the real world. In this paper, we present an approach to learn sensor fusion weights within the classification network in order to enable robust human activity recognition. In support of evaluating our method, we also introduce a novel case-study dataset recorded in a variety of conditions that specifically target the RGB, depth, and thermal camera modalities commonly available on robots. We show that our proposed approach is able to attain superior performance to approaches relying on individual modalities and is able to match or exceed the performance of multimodal methods in significantly less time.

References

[1] A. Wilson, K. Gupta, B. H. Koduru, A. Kumar, A. Jha, and L. R. Cenkeramaddi, "Recent advances in thermal imaging and its applications using machine learning: a review," *IEEE Sensors Journal*, 2023.

[2] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *International Journal of Robotics Research (IJRR)*, vol. 32, no. 8, pp. 951–970, 2013.

[3] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: a large-scale benchmark for 3d human activity understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 2684–2701, 2020.

[4] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *IEEE International Conference on Image Processing (ICIP)*, 2015.

[5] I.-C. Chen, C.-J. Wang, C.-K. Wen, and S.-J. Tzou, "Multi-person pose estimation using thermal images," *IEEE Access*, vol. 8, 2020.

[6] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, pp. 172–186, 2021.

[7] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A human activity recognition system using skeleton data from rgbd sensors," *Computational Intelligence and Neuroscience*, 2016.

[8] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3d skeletal data: A review," *Computer Vision and Image Understanding (CVIU)*, vol. 158, pp. 85–105, 2017.

[9] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient human pose estimation from single depth images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2821–2840, 2013.

[10] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2010.

[11] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "Movinets: mobile video networks for efficient video recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[12] M. Tan and Q. Le, "Efficientnet: rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning (ICML)*, 2019.

[13] A. Piergiovanni, A. Angelova, and M. S. Ryoo, "Tiny video networks," *Applied AI Letters*, vol. 3, no. 1, p. 38, 2022.

[14] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] B. Zhou, P. Wang, J. Wan, Y. Liang, and F. Wang, "A unified multimodal de- and re-coupling framework for rgb-d motion recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 11428–11442, 2023.

[16] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-ego: A large-scale dataset of paired third and first person videos," *arXiv preprint arXiv:1804.09626*, 2018.

[17] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about Kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018.

[18] M. Zia Uddin, W. Khaksar, and J. Torresen, "A thermal camera-based activity recognition using discriminant skeleton features and rnn," in *IEEE International Conference on Industrial Informatics*, 2019.

[19] J. Miura, M. Demura, K. Nishi, and S. Oishi, "Thermal comfort measurement using thermal-depth images for robotic monitoring," *Pattern Recognition Letters*, vol. 137, pp. 108–113, 2020.

[20] K. Song, Y. Zhao, L. Huang, Y. Yan, and Q. Meng, "RGB-T image analysis technology and application: a survey," *Engineering Applications of Artificial Intelligence*, vol. 120, p. 105919, 2023.

[21] S. Khalifa, G. Lan, M. Hassan, A. Seneviratne, and S. K. Das, "Harke: human activity recognition from kinetic energy harvesting data in wearable devices," *IEEE Transactions on Mobile Computing*, vol. 17, no. 6, pp. 1353–1368, 2018.

[22] F. Han, X. Yang, C. Reardon, Y. Zhang, and H. Zhang, "Simultaneous feature and body-part learning for real-time robot awareness of human behaviors," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[23] B. Reily, Q. Zhu, C. Reardon, and H. Zhang, "Simultaneous learning from human pose and object cues for real-time activity recognition," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[24] B. Reily, P. Gao, F. Han, H. Wang, and H. Zhang, "Real-time recognition of team behaviors by multisensory graph-embedded robot learning," *International Journal of Robotics Research (IJRR)*, 2022.

[25] S. Rho, S. M. Park, J. Pyo, M. Lee, M. Jin, and S.-C. Yu, "Lidar-stereo thermal sensor fusion for indoor disaster environment," *IEEE Sensors Journal*, vol. 23, no. 7, pp. 7816–7827, 2023.

[26] A. Gong, L. Huang, J. Shi, and C. Liu, "Unsupervised RGB-T saliency detection by node classification distance and sparse constrained graph learning," *Applied Intelligence*, vol. 52, no. 1, pp. 1030–1043, 2022.

[27] M.-H. Sheu, S. S. Morsalin, S.-H. Wang, L.-K. Wei, S.-C. Hsia, and C.-Y. Chang, "FHI-Unet: faster heterogeneous images semantic segmentation design and edge ai implementation for visible and thermal images processing," *IEEE Access*, vol. 10, 2022.

[28] J. U. Kim, S. Park, and Y. M. Ro, "Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, 2021.

[29] L. Jun, L. Zhongqiang, and X. Xingzhong, "RGB-T long-term tracking algorithm via local sampling and global proposals," *Signal, Image and Video Processing*, vol. 16, no. 8, pp. 2221–2229, 2022.

[30] Y. Miao, N. Huang, X. Ma, Q. Zhang, and J. Han, "On exploring pose estimation as an auxiliary learning task for visible–infrared person re-identification," *Neurocomputing*, vol. 556, p. 126652, 2023.

[31] A. J. Lee, Y. Cho, Y.-s. Shin, A. Kim, and H. Myung, "Vivid++: vision for visibility dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6282–6289, 2022.

[32] S. Vidas, P. Moghadam, and M. Bosse, "3D thermal mapping of building interiors using an RGB-D and thermal camera," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2013.

[33] R. Luo, O. Sener, and S. Savarese, "Scene semantic reconstruction from egocentric RGB-D-Thermal videos," in *IEEE International Conference on 3D Vision (3DV)*, 2017.

[34] M. O. Simón, C. Corneanu, K. Nasrollahi, O. Nikisins, S. Escalera, Y. Sun, H. Li, Z. Sun, T. B. Moeslund, and M. Greitans, "Improved RGB-D-T based face recognition," *IET Biometrics*, vol. 5, no. 4, pp. 297–303, 2016.

[35] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2012.

[36] S. Mahmud, M. Tonmoy, K. K. Bhaumik, A. M. Rahman, M. A. Amin, M. Shoyaib, M. A. H. Khan, and A. A. Ali, "Human activity recognition from wearable sensor data using self-attention," in *European Conference on Artificial Intelligence*, 2020.