Activity Recognition in Smart Homes with Knowledge Graph and Attention-Guided Learning

Sevakram T. Kumbhare, Ananda S. Chowdhury

Department of Electronics and Telecommunication Engineering
Jadavpur University, Kolkata-700032, India.
{stkumbhare.etce.rs, as.chowdhury}@jadavpuruniversity.in

Abstract. Activity recognition in smart homes has recently gained significant attention among the researchers in both computer vision and multimedia communities. Applications like video surveillance, and elderly care require a comprehensive understanding of different activities occurring in daily lives. However, this task remains relatively underexplored due to availability of only limited annotated data. In this paper, we introduce a model to first recognize fine-grained actions, and subsequently, identify composite activities resulting from such actions, in smart home environments. An Attention-based Shallow Residual Network (ASRNet) is applied first for feature extraction, where data scarcity is explicitly addressed by utilizing multiple datasets. We then perform fine-grained action recognition with an Aquila Hunger Games Search-optimized Bidirectional Long Short-Term Memory (AHS-BiLSTM) network. We then use the detected actions to identify composite activities. We construct an Activity Knowledge Graph (AKG), a weighted undirected graph, from the annotated data. Using a Bayesian inference on AKG, we identify composite activities from the already detected fine-grained actions. Experimental results on the "MPII Cooking 2" and "Toyota Smarthome" datasets demonstrate the effectiveness of our approach.

Keywords: Action Recognition · Composite Activity · Smart Homes · Knowledge Graph · Bayesian Inference.

1 Introduction

Activity recognition in smart homes [1] has drawn significant attention within the computer vision and multimedia communities. It has various applications, such as video surveillance, human-robot interaction, and elderly care. Understanding composite activities, is crucial for video surveillance and elderly care. As shown in Fig. 1, a composite activity "Breakfast" is composed of various fine-grained actions, like 'cut', 'drink', which may not always follow a specific temporal order [1, 2]. So, recognizing both fine-grained actions and high-level composite activities are essential [2]. However, due to limited annotated data, identifying composite

activities is difficult. Also, these activities can be performed differently, making it impractical to develop a visually annotated training set containing all possible variations.

(a)

(b)

Fig. 1. Composite Activities: (a) "Breakfast", and (b) "Cook"; each composite activity is shown with a set of constituent fine-grained actions.

Recently, some works [3] have been reported for activity recognition. Early methods focused on feature engineering, leveraging temporal information in videos. They make use of this knowledge to design specific hand-crafted representations [4]. For example, Rohrbach *et al.* [2] employed hand-centric and pose-based features to recognize fine-grained activities. They also utilized script data for composite activity recognition. It is the only work that reported results for composite activities in daily-life videos. However, they solely focus on the kitchen scenario. More recently, Das *et al.* [1] introduced a 3D CNN-based pose-guided spatio-temporal attention network for activity recognition in smart homes. This paper focuses on three scenarios, *i.e.*, kitchen, hall, and dining for fine-grained action recognition. Additionally, they discuss a composite and concurrent activity without reporting results for these tasks.

In this paper, we propose an attention-guided learning and knowledge graph to recognize fine-grained actions and high-level composite activities in smart homes. The proposed framework addresses the issue of data scarcity by training on multiple datasets. It consists of three main steps. First, an Attention-based Shallow Residual Network (ASRNet) extracts features from input videos. Next, an Aquila Hunger Games Search-optimized Bidirectional Long Short-Term Memory (AHS-BiLSTM) network is deployed to recognize fine-grained actions from the extracted features. Finally, we identify composite activities using detected actions. We construct an Activity Knowledge Graph (AKG) from annotated

data [5]. Bayesian inference [6] is applied on this graph to infer about a composite activity from the already detected actions. Below, we summarize the key contributions of this work.

1. We present an Attention-based Shallow Residual Network (ASRNet) for fine-grained action recognition in smart homes using multiple datasets. The goal is to achieve accurate action recognition while minimizing overall network parameters.
2. We apply Bayesian inference on an Activity Knowledge Graph (AKG), to effectively identify and analyze composite activities from already detected actions. This structured approach enhances the generalizability of our solution in handling the intricacies of composite activity recognition.
3. We introduce an Aquila Hunger Games Search (AHS) algorithm to optimize BiLSTM network parameters. This makes our solution more computationally efficient.

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 provides details of the proposed model. Section 4 presents experimental results with ablation studies. Finally, Section 5 concludes the paper and outlines future research directions.

2 Related Work

Human activity recognition is crucial for different applications in computer vision and multimedia [1, 7, 8]. Among the various challenges in this domain, the recognition of composite activities stands out for its complexity and real-world relevance. Historically, approaches to activity recognition have relied on local features such as dense trajectories [9] and Fisher vector encoding [10]. These approaches are simple and effective for small datasets. For larger datasets, researchers typically integrate local features with those learned by convolutional networks. However, deep-learning techniques like Two-stream ConvNets [11] struggled to encode long-range temporal information. To address this, Donahue *et al.* [12] extracted spatial features from a CNN network and fed them into sequential networks such as LSTM. I3D [13] further advanced the field by inflating the kernels of ImageNet pre-trained 2D CNNs to jump-start 3D CNN training. Sahu *et al.* [14] utilize graphs for egocentric action recognition. Despite their effectiveness in recognizing fine-grained and object-based activities, these methods are often found to be computationally expensive.

In the context of sensor data, we can find some approaches [3, 15] for composite activity recognition. For instance, Thapa *et al.*[15] developed a hybrid technique for complex activity recognition based on Skip-Chain Conditional Random Field (SCCRF) and Bi-directional Long-Short Term Memory (BiLSTM). However, the recognition of composite activities in video data has received little attention in the existing literature. Rohrbach *et al.* [2] introduced the "MPII Cooking 2" dataset for recognizing fine-grained actions and composite activities. They make use of hand-centric and pose-based features, as well as script data.

More recently, Das *et al.* [1] presented the "Toyota Smarthome" dataset, as well as a 3D CNN-based pose-guided spatio-temporal attention network for activity recognition in smart homes. This paper discusses composite and concurrent activities, though they did not report results for these activities. In summary, the existing approaches fail to address the challenges of composite activity recognition in video data from smart homes. Also, the existing models are not fully optimized for real-world applications.

In contrast, we present an attention-guided learning and knowledge graph to address the problem of recognizing fine-grained actions, as well as high-level composite activities in smart homes. Our framework is trained on multiple datasets to mitigate data scarcity.

3 Proposed Framework

Our solution pipeline includes three key steps: feature extraction, fine-grained action recognition, and composite activity recognition. Fig. 2 provides a block diagram of our proposed framework. A detailed description of each of the components is given below.

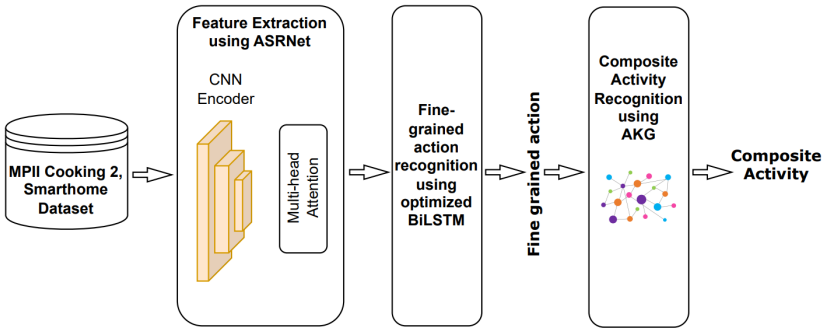

Fig. 2. A Multi-dataset composite activity recognition framework

3.1 Feature extraction

In this work, we introduce an Attention-based Shallow Residual Network (ASRNet) designed for feature extraction. The aim is to capture and refine video features for action recognition while minimizing overall network parameters. The ASRNet integrates a residual structure [16] with an attention mechanism [17] to improve fine-grained action recognition performance. The architecture of ASRNet is given in Fig. 3. During training, the network takes input from multiple

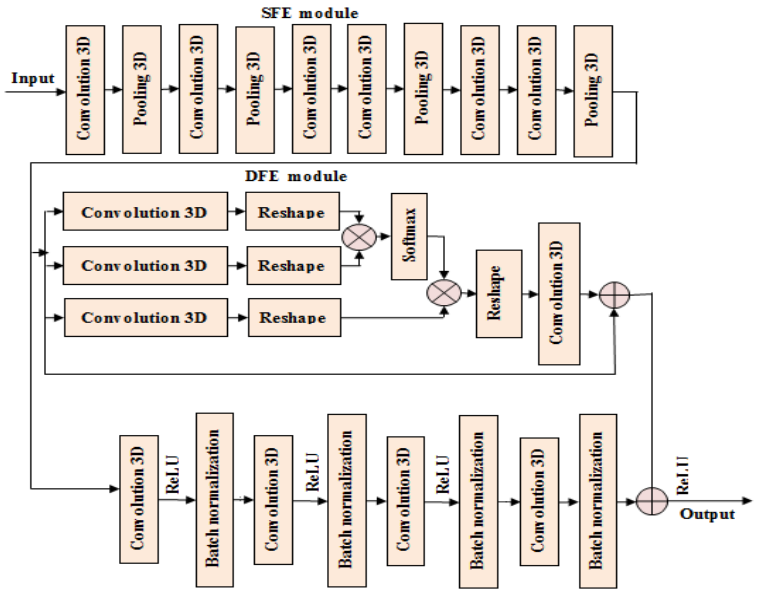

Fig. 3. Attention-based Shallow Residual Network (ASRNet)

datasets. We create mixed batches from these datasets to allow the network to learn from diverse data in each iteration.

ASRNet consists of a Shallow Feature Extractor (SFE) and an attention-based Deep Feature Extractor (DFE) module. The SFE module enhances the 3D residual structure. It processes multi-frame sequences as input. It includes four smaller blocks, each with 3D convolution and pooling operations. These layers capture initial spatial and temporal features from the video frames. The output from the SFE module is then passed to the attention-based DFE module. The DFE employs multi-head attention [17] along with a skip connection comprising a series of convolution and batch normalization layers. The multi-head attention mechanism captures global features. It preserves the key information of video frames. The output from the attention block is combined with the skip connection via ReLU activation, producing the final extracted features.

3.2 Fine-grained action recognition

After feature extraction, the AHS-BiLSTM network is applied for fine-grained action recognition. It leverages the capabilities of a Bidirectional Long Short-Term Memory (Bi-LSTM) network [12]. The LSTM classifier consists of four main components namely input gate, memory cell, output gate and forget gate. The memory cell retains data for varying durations, either long or short. It facilitates the model's ability to remember important information. The input gate

regulates the amount of data entering the memory cell. The forget gate is utilized for controlling the details of LSTM cell. The output gate computes the output activation based on the cell's information. Bi-LSTM is developed to resolve the issue related to exploding and vanishing gradient using gated cells. The structure of Bi-LSTM is shown in Fig. 4. The AHS algorithm further optimizes the Bi-LSTM model's performance through hyperparameter tuning. The AHS-BiLSTM framework ensures robust and precise action recognition by effectively managing temporal dependencies and improving classification accuracy through enhanced feature representation and optimization.

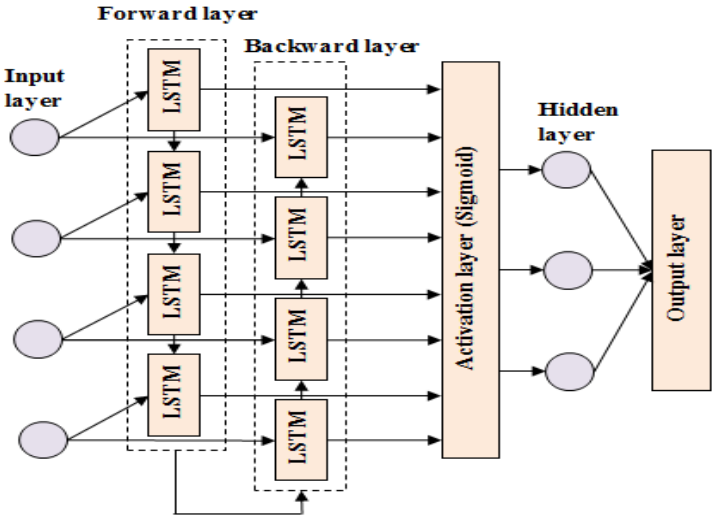

Fig. 4. Bi-LSTM structure

AHS optimization The Aquila Hunger games Search (AHS) algorithm is used to tune the hyperparameters of the Bi-LSTM model. AHS is a population-based, gradient-free optimization algorithm that imitates the collaborative foraging behavior of social animals. The behaviors of social animals are based on the degree of hunger. This adaptive, weight-based approach simulates the hunger effect through logical rules applied at each step. It provides a simple and dynamic framework with high performance. AHS combines the principles of the Aquila optimizer [18] for initialization, capturing prey efficiently by soaring and gliding techniques. Hunger search optimization [19] then updates the positions of agents based on hunger-driven behaviors, influencing decision-making and movement. The Aquila optimizer soaring strategy allows it to survey a wide area before a rapid descent to capture prey, while the hunger search optimization uses hunger

as a driving factor for agent actions. The total number of candidate solutions is represented as follows.

$$Y_j = r * (U_j - L_j) + L_j; \quad j \in [1, P] \tag{1}$$

where r is a random number, j is the agent in population P, L_j represents the j^{th} lower bound, U_j is the j^{th} upper bound.

Approaching food In each iteration of the AHS algorithm, the location of each agent is updated based on the collaborative and independent behaviors of the population [19]. Some agents hunt in groups, while others act alone. The behavior of approaching food is represented by Equation 2. It ensures diverse potential solutions in the search space, enhancing the exploration capabilities of the algorithm.

$$Y(u+1) = \begin{cases} Y(u) \cdot (1 + r(1)) & , s_1 < m \\ X_1 \cdot Y_c(u) + S \cdot X_2 \cdot |Y_c(u) - Y(u)| & , s_1 > m, s_2 > F \\ X_1 \cdot Y_c(u) - S \cdot X_2 \cdot |Y_c(u) - Y(u)| & , s_1 > m, s_2 < F \end{cases} \tag{2}$$

where u represents the current iteration, $Y(u)$ indicates the location of an individual, $Y_c(u)$ indicates the location of the best individual in the current iteration, $r(1)$ indicates the random number with normal distribution, $Y(u) \cdot (1 + r(1))$ reflects an agent's food searching ability combining both hunger and randomness at its present location, X_1, X_2 are the hunger weights, m is the enhancement parameter used in this approach, $s_1, s_2 \in [0, 1]$ are the random numbers, $Y_c(u) - Y(u)$ indicates the range of activity for individual, F refers to variation control, and $S \in [-b, b]$. The value of b is determined based on current and maximum number of iterations. These parameters are used to calculate the new positions of the agents.

Hunger strategy The hunger strategy in AHS dynamically adjusts the hunger weights X1 and X2 for each agent, as shown in Equations 3 and 4.

$$X_1(j) = \begin{cases} hny(j) \cdot \frac{P}{THny} \cdot s_3 & ; s_3 < m \\ 1 & ; otherwise \end{cases} \tag{3}$$

$$X_2(j) = 2 \cdot s_4 \cdot (1 - exp(-|hny(j) - THny|)) \tag{4}$$

where, the hunger value of j^{th} agent is indicated with $hny(j)$, P is the number of agents, $THny$ represents the sum of hunger values of all agents, $s_3, s_4 \in [0, 1]$ are the random number. The hunger value, $hny(j)$ is given as in Equation 5.

$$hny(j) = \begin{cases} 0 & ; G(j) = G_{bst} \\ hny(j) + H & ; G(j) \neq G_{bst} \end{cases} \tag{5}$$

where, $G(j)$ represents the fitness value of each individual, and G_{bst} is the highest fitness value achieved in the current iteration. In each iteration, the hunger

value of the best-performing agent is set to zero, ($hny(Y_c) = 0$). For other agents, a new hunger value H is added based on their initial hunger. It promotes exploration and prevents premature convergence. This strategy introduces new sensations of hunger, adding diversity to the search process. The algorithm also considers search space boundaries and fitness values of agents to determine their foraging capacity and food demand. The proposed AHS optimization algorithm is simple and efficient, suitable for both constrained and unconstrained continuous problems.

3.3 Composite activity recognition

We construct an activity knowledge graph $AKG = (V, E)$ for composite activity recognition. It encapsulates the relationships between composite activities and their fine-grained actions. This approach leverages the inherent structure and dependencies of human activities, providing a robust framework for real-world applications. In the graph, V represents the set of nodes, where each node $v_i \in V$ corresponds to composite activity label, and node $v_j \in V$ represents a fine-grained action label. The edges $(v_i, v_j) \in E$ capture the contextual relationships between action and composite activity. Formally, the edge set E is defined as shown in Equation 6.

$$E = \{(v_i, v_j) \mid v_i, v_j \in V\} \tag{6}$$

To quantify the likelihood of action related to a composite activity, we assign probabilities to the edges. Let $P(v_j|v_i)$ represent the probability that the action v_j related to composite activity v_i as shown in Equation 7. These probabilities are estimated from the annotated training data.

$$P(v_j|v_i) = \frac{N(v_i, v_j)}{N(v_i)} \tag{7}$$

Here, $N(v_i, v_j)$ denotes the number of times action v_j contribute to composite activity v_i, and $N(v_i)$ denotes the total occurrences of composite activity v_i. For computational efficiency, we represent the knowledge graph using an adjacency matrix A as defined in Equation 8.

$$A_{ij} = P(v_j|v_i) \tag{8}$$

Fig. 5 illustrates a representative knowledge graph for the composite activities "Cook" and "Make_coffee", along with their associated actions. It highlights the relationships between composite activity and related actions. It shows that a single action can be linked to multiple composite activities.

We apply Bayesian inference on the knowledge graph to predict the composite activity based on the set of already detected actions. The goal is to predict the composite activity C that best matches the observed sequence of actions. For this, we consider T consecutive predicted actions, where T is determined by the dataset training annotations available, to ensure optimal recognition accuracy. In our approach, T is set to 8. Given a sequence of recognized fine-grained actions

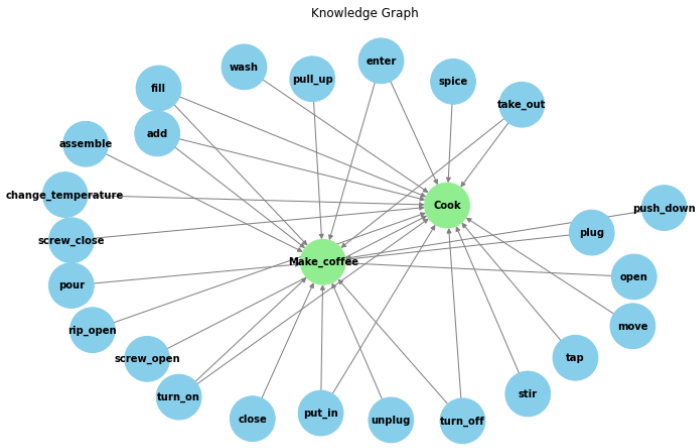

Fig. 5. Representative knowledge graph for composite activities "Cook" and "Make_coffee", along with their associated actions

$\mathcal{X} = \{x_1, x_2, \cdots, x_T\}$, we compute the likelihood of each composite activity C. To compute $P(C|\mathcal{X})$, we use the probabilities encoded in the knowledge graph. Each edge (v_i, v_j) has an associated probability $P(x_j|C)$, estimated from training data. For a given composite activity C with fine-grained actions, the likelihood of C given the observed sequence of fine-grained actions $\mathcal{X}$, is computed as shown in Equation 9.

$$P(C \mid \mathcal{X}) \propto \left(\prod_{j=1}^{T} P(x_j \mid C) \right) \cdot P(C) \tag{9}$$

Here, $P(x_j \mid C)$ is the probability of observing action x_j given the composite activity C. We assume uniform distribution for composite activities, *i.e.* $P(C) = 1/k$, where k is the number of composite activities. To infer the most likely composite activity $\hat{C}$ based on the observed sequence X, maximize the likelihood $P(C \mid X)$ as shown in Equation 10.

$$\hat{C} = \arg\max_{C \in \mathcal{C}} P(C \mid \mathcal{X}) \tag{10}$$

where $\mathcal{C}$ is the set of all possible composite activities.

4 Experimental results

In this section, we present the experimental results. All experiments are carried out on a desktop PC with Intel(R) Core(TM) i5-12400F @ 2.50GHz, 32GB DDR5 RAM and NVIDIA RTX 12GB GPU. We execute tests on the benchmark datasets MPII Cooking 2 [2] and Toyota Smarthome [1]. We present results

for fine-grained action recognition on both datasets. However, we compare our composite activity recognition results only on the MPII Cooking 2 dataset, as it is the only dataset with previously reported results for this task. We report mean average precision (mAP) [2] for the MPII Cooking 2 dataset and mean per class accuracy [1] for the Toyota Smarthome dataset.

4.1 Dataset

MPII Cooking 2 The dataset contains videos of cooking activities. Rohrbach *et al.* [2] presented the MPII Cooking 2 dataset to address the issue of fine-grained actions and composite activities in a kitchen scenario. The dataset includes 273 videos. It contains fine-grained actions, such as, 'cut dice', 'take lid', and 'put lid', and so on. Additionally, it has 59 composite activities. The number of actors is 30. The dataset is separated into train and test groups according to subject. The train set includes 24 videos of different subjects, with the remaining 6 used for testing.

Toyota Smarthome The Toyota Smarthome dataset [1] consists of real-world activities performed by humans in their daily lives, such as reading, breakfast, and so on. The dataset is collected across three different scenes: the dining room, living room, and kitchen. The dataset includes 16,115 videos of 31 action classes. The videos are captured from 7 different camera perspectives. It also contains 3 composite activities. The videos have resolution of 640 $\times$ 480 and offers 3 modalities: RGB, Depth and 3D skeleton. For privacy, the face of the subjects is blurred using tinyface detection method.

4.2 Results for fine-grained action recognition

MPII Cooking 2 dataset In this section, we present results of fine-grained action recognition on the MPII Cooking 2 dataset. Table 1 presents how well our method worked in fine-grained action recognition in MPII Cooking 2 dataset. We compare our technique with recent approaches, like three-stream [20], TSN [21], region sequence CNN [22], TSM [23] and AVR [24]. Our ASRNet framework achieves best mean average precision (mAP) of 75.3%, outperforming all state-of-the-art methods.

Toyota Smarthome dataset Table 2 presents the results for fine-grained action recognition on Toyota Smarthome dataset. We compare our technique with recent approaches, such as, DT [9], I3D [13], Separable STA [1], and Action Genome [25]. Our ASRNet framework achieves best mean per class accuracy of 64.9%. We outperform all state-of-the-art methods.

During inference, it is observed that the AHS algorithm used to optimize the BiLSTM network reduces inference time for fine-grained action recognition by 18%.

Table 1. Comparison with state-of-the-arts methods for fine-grained action recognition on MPII Cooking 2 dataset

Method	**mAP %**
Dense Trajectories [2]	34.5
Three-stream [20]	55.6
TSN [21]	68.5
Region-sequence CNN [22]	70.3
TSM [23]	71.2
AVR [24]	73.7
Ours	**75.3**

Table 2. Comparison with state-of-the-art methods for fine-grained action recognition on Toyota Smarthome dataset

Method	**Mean per class %**
DT [9]	41.9
I3D [13]	53.4
Separable STA [1]	54.2
Action Genome [25]	63.6
Ours	**64.9**

4.3 Results for composite activity recognition

This section presents the results for composite activity recognition on MPII Cooking 2 dataset. We train on 126 videos corresponding to the 31 test categories. Table 3 presents the results for composite activity recognition on MPII Cooking 2 dataset. Rohrbach *et al.* [2] is the only other work that reported results for composite activities in daily-life videos. Fig. 6 provides a qualitative

Table 3. Comparison with state-of-the-art methods for composite activity recognition on MPII Cooking 2 dataset

Method	**mAP %**
Dense Trajectories + SVM [2]	39.8
Dense Trajectories + Hand-Trajectories + SVM [2]	41.1
Dense Trajectories + Hand-Trajectories + SVM + Auto-Segment[2]	56.9
Ours	**59.2**

result of the composite activity "Cook" alongside its ground truth. The frames are representative of their respective actions. Composite activity recognition is constrained by the limited availability of annotated datasets.

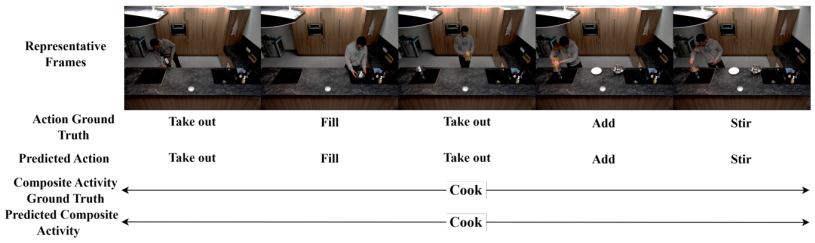

Fig. 6. Qualitative results for composite activity recognition along with its ground truth

4.4 Ablation Study

In this section, we investigate the benefits of attention mechanism in the proposed ASRNet framework for fine-grained action recognition on MPII Cooking 2 dataset. We evaluated our pipeline without attention, using self-attention, and with multi-head attention mechanisms. Table 4 demonstrates that incorporating attention improves performance, with the multi-head attention mechanism yielding the best results.

Table 4. Effect of attention mechanisms on fine-grained action recognition in MPII Cooking 2 dataset

Method	**mAP %**
ASRNet (without attention)	72.5
With Attention [26]	74.1
With Multi-head Attention [17]	**75.3**

5 Conclusion

In this paper, we propose knowledge graph and attention-guided learning framework for recognizing fine-grained actions and composite activities in smart homes. We address data scarcity by conducting experiments across multiple datasets. Our approach involves feature extraction using the Attention-based Shallow Residual Network (ASRNet). Further, We introduce AHS to optimize BiLSTM network parameters for fine-grained action recognition. Finally, we detect composite activities by applying a Bayesian inference on an activity knowledge graph, built from the fine-grained action labels. Experimental comparisons with several recent methods on multiple datasets indicate the efficacy of our proposed solution. Future work will focus on improving activity recognition in untrimmed

videos. We will also develop a comprehensive model for composite activity recognition and evaluate its performance in real-time processing [27].

Acknowledgements I would like to express my sincere gratitude to Maharashtra Government (Government Polytechnic, Murtizapur) for the support and opportunity to pursue my Ph.D. studies through deputation. This invaluable support has enabled me to advance my research and deepen my expertise.

References

1. Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome: Real-world activities of daily living. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 833–842 (2019)
2. Rohrbach, M., Rohrbach, A., Regneri, M., Amin, S., Andriluka, M., Pinkal, M., Schiele, B.: Recognizing fine-grained and composite activities using hand-centric features and script data. Int. J. Comput. Vis. 119, 346—-373 (2016)
3. Dang, L.M., Min, K., Wang, H., Piran, M.J., Lee, C.H., Moon, H.: Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognit. 108, 107561 (2020)
4. Zheng, Y.D., Liu, Z., Lu, T., Wang, L.: Dynamic sampling networks for efficient action recognition in videos. IEEE Trans. Image Process. 29, 7970–7983 (2020)
5. Gao, J., Zhang, T., Xu, C.: I know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In: Proc. AAAI Conf. Artif. Intell., pp. 8303-8311 (2019).
6. Liu, C., Ying, J., Han, F., Ruan, M.: Abnormal Human Activity Recognition using Bayes Classifier and Convolutional Neural Network. In: Proc. IEEE 3rd Int. Conf. Signal and Image Process. (ICSIP) pp. 33-37 (2018).
7. Kumbhare, S., Chowdhury, A.: Learning from Multiple Datasets for Recognizing Human Actions. In: Proc. 13th Indian Conf. Comput. Vis. Graphics Image Proc. (ICVGIP), pp. 1-9, (2022)
8. Sahu, A., Chowdhury, A.: Together Recognizing, Localizing and Summarizing Actions in Egocentric Videos. IEEE Trans. Image Process.30, 4330-4340(Apr 2021)
9. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 3169–3176 (2011)
10. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 143–156 (2010)
11. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. Adv. Neural Inf. Process. Syst. 27, 568—-576 (Dec 2014)
12. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Darrell, T., Saenko, K.: Long-term recurrent convolutional networks for visual recognition and description. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2625–2634 (2015)
13. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 6299–6308 (2017)

14. Sahu, A., Bhattacharya, R., Bhura, P., Chowdhury, A. S.: Action recognition from egocentric videos using random walks. In: Proc. 3rd Int. Conf. Comput. Vis. Image Process. (CVIP), Volume 2, pp. 389-402 (2020)
15. Thapa, K., Abdullah Al, Z.M., Lamichhane, B., Yang, S.H.: A deep machine learning method for concurrent and interleaved human activity recognition. Sensors 20, 5770 (2020)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp.770-778 (2016)
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proc. 31st Int. Conf. Neural Inf. Process. Syst. p. 6000–6010 (2017)
18. Abualigah, L., Yousri, D., Abd Elaziz, M., Ewees, A.A., Al-Qaness, M.A., Gandomi, A.H.: Aquila optimizer: a novel meta-heuristic optimization algorithm. Computers & Industrial Engineering 157, 107250 (2021)
19. Mehta, P., Yildiz, B.S., Sait, S.M., Yildiz, A.R.: Hunger games search algorithm for global optimization of engineering design problems. Materials Testing 64(4), 524–532 (2022)
20. Wang, L., Huo, G., Li, R., Liang, P.: A descriptive behavior intention inference framework using spatio-temporal semantic features for human–robot real-time interaction. Engineering Applications of Artificial Intelligence 128, 107488 (2024)
21. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: Proc. Eur. Conf. Comput. Vis. (ECCV) pp. 20–36 (2016)
22. Ma, M., Marturi, N., Li, Y., Leonardis, A., Stolkin, R.: Region-sequence based six-stream cnn features for general and fine-grained human action recognition in videos. Pattern Recognit. 76, 506–521 (2018)
23. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 7083–7093 (2019)
24. Yenduri, S., Perveen, N., Chalavadi, V., Mohan, C.K.: Fine-grained action recognition using attribute vectors. In: VISIGRAPP (5: VISAPP). pp. 134–143 (2022)
25. Rai, N., Chen, H., Ji, J., Desai, R., Kozuka, K., Ishizaka, S., Adeli, E., Niebles, J.C.: Home action genome: Cooperative compositional action understanding. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 11184–11193 (2021)
26. Lin, Z., Feng, M., dos Santos, C.N., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A Structured self-attentive sentence embedding. In: Proc. Int. Conf. Learn. Represent. (ICLR) (2017)
27. Dai, R., Das, S., Sharma, S., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. IEEE Trans. Pattern Anal. Mach. Intell. 45(2), 2533–2550 (Feb 2023)