Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 246 (2024) 4104–4113

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Indoor human activity recognition based on context relationships

Tong Tong[a,*], Rossitza Setchi[a], Yulia Hicks[a]

[a]*Research Centre in AI, Robotics and Human-Machine Systems (IROHMS), Cardiff University, Cardiff CF24 3AA, UK*

Abstract

Human activity recognition, as a significant branch of artificial intelligence, requires increasingly generalized and precise methodologies due to growing demands. Therefore, this paper proposes a context-based method for recognising indoor human activities, which interlinks indoor human activities with interactions with objects, making the contextual relationship between humans and objects particularly crucial. In addition, this research has developed a new dynamic graph dataset based on publicly available video datasets and their associated descriptive scripts, instantiating the relationships between humans and objects. A novel architecture for human activity recognition is developed in this research. This architecture utilizes graph neural networks and self-attention mechanisms to learn the significance of the interactions between humans and objects and capture the relationships between video frames on a temporal level. The results demonstrate that the classification accuracy reaches 0.86 and it also performs better than other current advanced algorithms STGAT and STGCN. It is noteworthy that the approach also effectively reduces ambiguity in activity recognition.

Keywords: Human activity, Context relationship, Graph neural network, Scene graph

1. Introduction

Human Activity Recognition (HAR) represents a significant field of research that focuses on identifying and classifying various human behaviours or actions based on data derived from sensors or video recordings. This interdisciplinary domain intersects with computer science, data analysis, artificial intelligence (AI), and healthcare, among others, aiming to automate the process of understanding human activities through machine learning models

* Corresponding author.
E-mail address: tongt6@cardiff.ac.uk

1877-0509
10.1016/j.procs.2024.09.249

and other technologies [1]. HAR finds its application spreading across various sectors including healthcare, smart homes, surveillance, sports, and human-computer interaction.

As a research area HAR has progressed from using RGB or grayscale videos as input [2] due to their popularity and ease of access, to extracting both static and dynamic features from video data, captured through convolution and cross-correlation operators [3]. In addition, through wearable computing and ambient intelligence, where the primary focus was on enhancing the interaction between humans and technology [4], and with advancements in sensor technology and the proliferation of mobile devices equipped with a myriad of sensors such as accelerometers, gyroscopes, and GPS, the capacity to capture and analyse human movement and behaviour has significantly expanded [5]. Concurrently, the evolution of AI and machine learning techniques has propelled HAR into new heights, enabling the extraction of meaningful patterns from complex, multidimensional data generated by these sensors.

In healthcare, HAR is transforming patient care and monitoring, facilitating remote health monitoring systems that can predict and prevent adverse events by recognizing activities, such as falls among the elderly [6]. HAR is instrumental in developing intelligent systems that adapt to the inhabitants' behaviours, enhancing energy efficiency, security, and comfort. For instance, distinguishing when a person is sleeping or away from home allows for the adjustment of lighting and heating, thereby conserving energy. HAR contributes to public safety and security by detecting unusual or suspicious activities in real-time. In the context of context-aware human activity recognition (HAR), Claudio et al. suggested utilizing domain knowledge related to high-level contextual data to enhance the accuracy of predictions made by an activity classifier [7].

Despite the progress, HAR faces several challenges, including the diversity of human activities, and the need for large annotated datasets for training robust models [8]. In the HAR field based on traditional vision and sensor methods, the complexity of the environment can cause inaccurate recognition results. Successful identification of such activities as pouring water, eating food, plays a vital role in caring for the elderly and monitoring hazardous activities. However, most of the current research focuses on individual activities such as running and jumping. The actual research on indoor activity recognition is still insufficient.

Context is very important for identifying indoor activities as most of them are accompanied by interactions with objects [9]. For example, when a person drinks water, the person may pick up a kettle, pick up a water cup, pour water in a series of interactive behaviours with objects. Therefore, when identifying indoor activities, the system should focus on the surrounding context.

The focus of this paper is to explore recognition of indoor human activities by identifying people and the surrounding objects, and determining the relationship between people and objects, so as to identify current human activities. In reality, it is difficult to identify different activities indoors by relying only on sensors such as gyroscopes. Therefore, this paper uses computer vision techniques in combination with context relationships between people and objects as clues to identify human activities. It is hypothesised when a person performs activities indoors, they will interact with the surrounding context, and human activities can be identified based on these clues.

2. Related work

The field of Human Activity Recognition (HAR) has witnessed substantial growth and diversification, reflecting the advancements in sensor technologies, machine learning algorithms, and their applications across various domains. This section explores the related work in HAR, highlighting significant contributions, methodologies, and emerging trends that shape the current landscape of this research area.

2.1. Human activity recognition

Ji et al. [10] utilized a human detector to segment humans in videos, and then input the segmented videos into a novel 3D CNN model to extract spatiotemporal features from the videos. Skeleton sequences represent informative human motion. Wang et al. [11] proposed skeleton spectra and joint trajectory maps. They all encode the spatiotemporal information of the skeleton sequence into colour texture images, and then use CNN to perform HAR. HAR relied on wearable sensors due to their effectiveness in capturing body movements. For instance, Fullerton et al. [12] demonstrated the use of accelerometers worn on the body to classify daily activities with notable accuracy.

The application of deep learning in HAR signifies a paradigm shift, offering the ability to automatically learn feature representations from raw data, thus bypassing the labour-intensive process of manual feature extraction. Sharma et al. [13] applied Convolutional Neural Networks (CNNs) for activity recognition using wearable sensor data, showcasing superior performance compared to traditional machine learning approaches. Following this, Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks, have been extensively adopted for modelling time-series sensor data, given their proficiency in capturing temporal dependencies [14].

The integration of Attention Mechanisms with RNNs further enhanced the model's ability to focus on salient features in the data sequence, leading to improved accuracy in complex activity recognition tasks [15]. Moreover, the emergence of Graph Convolutional Networks (GCNs) has opened new avenues for HAR by exploiting the relational information among skeleton, offering a more holistic view of human activities [16].

The methodology of HAR involves several stages, starting with data acquisition through sensors. The raw data, characterized by its high dimensionality and variability, undergoes preprocessing to filter noise and extract relevant features. Following this, various machine learning models—ranging from traditional approaches like decision trees and support vector machines to advanced techniques such as deep learning—are employed to classify the activities based on the extracted features. Deep learning, in particular, has shown remarkable success in HAR due to its ability to learn complex representations of data.

2.2. Scene graph generation

In recent years, the integration of scene graph generation with human activity recognition has garnered significant attention in the field of computer vision. Scene graphs provide structured representations of scenes by detailing objects and their interrelationships. They are particularly useful for comprehensively understanding dynamic human activities in various contexts. The work in this domain is by Chang et al. [17], who introduced a method for generating scene graphs from images and then applying these graphs to enhance video understanding tasks. Fabian suggested that efficient context cascading is very effective for fast human activity recognition [18].

He et al. [19] further explored this idea by proposing a novel framework that leverages scene graph generation to detect and describe human-object interactions. By capturing these interactions as edges in a scene graph, their model was able to achieve more nuanced activity recognition, particularly in complex scenes with multiple actors and interactions. Moreover, the work by Yang et al [20]. efficiently deals with the quadratic number of potential relations between objects in an image. This method effectively captures contextual information between objects and relations.

Research by Ost et al. [21] proposes the first neural rendering method that represents multi-object dynamic scenes as scene graphs. Their approach introduces a learned scene graph representation that encodes object transformations and radiance. It has been demonstrated that can be learned within dynamic scenes.

2.3. Graph neural network

Graph neural networks (GNNs) have emerged as a powerful tool in the field of human activity recognition, offering significant advancements in understanding complex interactions within data. The application of GNNs in this domain leverages their ability to model relationships and dependencies between objects and humans in various environments, enhancing the accuracy and robustness of activity recognition systems.

One key study by Yan et al. proposed a new dynamic skeleton model called Spatio-Temporal Graph Convolutional Network (ST-GCN). This model overcomes the limitations of previous methods by automatically learning spatial and temporal patterns from data [22]. Heidari and Alexandros [23] analysed the spatial attention used in the Spatio-Temporal GCN layers and proposed a symmetric spatial attention mechanism to better reflect the symmetrical characteristics of human joint positions during the execution of actions.

In the field of video-based pedestrian re-identification, modelling the temporal relationships between different frames and the spatial relationships within frames may potentially solve issues related to occlusions by visually similar negative samples and visual blurriness [24]. The level of engagement of a driver during driving plays a crucial role and can lead to negative outcomes for driving safety. Pan et al. [25] utilise a Graph Convolutional Network (GCN) to infer spatial structural features within a single frame, followed by a Long Short-Term Memory (LSTM) network for

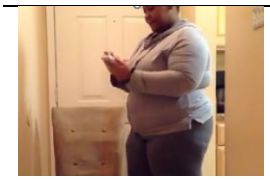
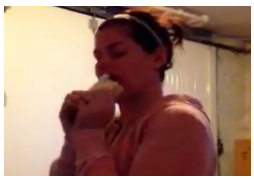
learning temporal motion features within sequences. This approach has successfully identified eight common activities related to driving.

These studies collectively highlight the growing importance of graph-based approaches in capturing the complex and dynamic nature of human activities, paving the way for more intuitive and effective recognition systems. The continuous evolution of GNN architectures and their applications in this field underscores the potential for even greater advancements in the near future.

3. Dataset

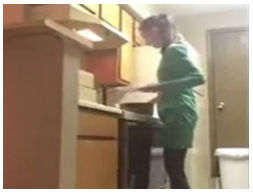
As previously mentioned, for the recognition of human activities indoors, it is essential to focus on the relationship between humans and the objects around them. In other words, context information is very important. Currently, there is no dataset available that provides the aforementioned information. Therefore, this paper has created a new dataset that has evolved from existing video data. The original video data used in this paper is from Action Genome [26], which annotates the relationships between people and objects within the videos. Moreover, it also provides scripts describing the activities or behaviours in each video. However, Action Genome is not intended as a dataset for human activity recognition, but each video is closely related to human activities. Therefore, this paper uses Action Genome as the initial dataset. It was observed that not every video clearly shows human activity, and the scripts do not always accurately describe whether clear human activity is present. The human activities selected for this paper are those that persist over a period of time in the videos. In other words, mere momentary actions are not included in the data annotation as human activities. This study first filtered the original video scripts and selected 6 common activities in current human activity recognition research based on related words, such as water, cooking, food, etc. Since the original video data does not have clear human activity labels and unified paradigm for original video scripts, this study used ChatGPT to analyse the scripts of the selected videos, deriving corresponding human activities labels based on the video scripts, which were then used as training labels. This study also conducted human observations on each of them to ensure that the tags generated by ChatGPT were rigorous and correct. Finally, 6 types of human activities were selected for recognition, and 150 videos totally. Table 1 shows a frame from each video representing the activities as examples. The generated activity labels were subsequently checked for accuracy by one of the authors of this paper.

Table 1. Examples of frames representing different activities.

A frame from the video	Scripts	Human activity
	The person gets distracted by their phone, and begins playing with it.	Playing phone
	A person is seen grasping the food eating it.	Eating food
	A person takes a broom from the corner and begins working on tidying up.	Tidying up

A person is seen sitting down at a table. They begin pouring themselves some water. Pouring water

A person is seen smiling and watching television. Watching television



A person turns the light on over the stove and begins cooking. Cooking

Once we have the final data, it is necessary to instantiate the humans, objects, and their relationships. As previously stated, graph data can store this type of data well and can be beneficial for subsequent modelling. Therefore, this study uses a scene graph generation algorithm based on the STTran [27]. This algorithm employs a faster R-CNN to identify objects in videos and uses a self-attention mechanism to capture the relationships between frames of the video.

The input X, which is represented by each frame, $X \in R^{(N\times D)}$ which has N entries of D dimensions, is transformed into queries ($Q = XW\_Q, W\_Q \in R^{(N\times D\_q)}$), keys ($K = XW\_K, W\_K \in R^{(N\times D\_k)}$) and values ($V = XW\_V, W\_V \in R^{(N\times D\_v)}$) through linear transformations. W_Q, W_K, W_V are trainable parameter matrices which add the input to get the Q, K, V. Every value will get through the attention defined by:

$$Attention\ (Q,K,V) = softmax(\frac{QK^T}{\sqrt{D_k}}) \tag{1}$$

The next step involves using Faster R-CNN to extract relationship representation for each frame in a given video at time t. The spatial information and semantic embeddings are formulated as:

$$x_t^k = \langle (W_s\ v_t^i, W_o\ v_t^i, W_{u\varphi}\ (u_t^{ij} \oplus f_{box}\ (b_t^i, b_t^j))), s_t^i, s_t^j \rangle \tag{2}$$

Where $\langle . \rangle$ is concatenation operation, φ is flattening operation and $\oplus$ is addition. W_s, W_o, and W_u represent the linear matrices for dimension compression, u_t^{ij} indicates the feature map of the union box, while f_{box} is the function transforming the bounding boxes of subject and object to an entire feature. The semantic embedding vectors s_t^i, s_t^j are determined by the object categories of subject and object.

Finally, a spatial encoder, frame encoder and temporal decoder are used to generate the dynamic graph. The frame encoder concatenates every frame, while the temporal decoder classifies each frame and generates the scene graph:

$$X_t^{(n)} = Att_{enc.}\left(Q = K = V = X_t^{(n-1)}\right) \tag{3}$$

Where X_t is a single input after spatial encoding, and the queries Q, keys K, values V share the same input and output of n encoder layer. To be more specific, $X_t^{(n)}$ is the output of $X_t^{(n-1)}$ through the attention encoding.

To capture the temporal dependencies between video frames, it employs a sliding window approach with a parameter η to batch the frames together. This method ensures that only adjacent frames interact, minimizing the inclusion of irrelevant

information. The contextualized representations $[X1, \ldots, XT\ ]$ and the $i-th$ generated input batch is presented as:

$$Z_i = [X_i, \ldots, X_{i+\eta-1}], i \in \{1, \ldots, T - \eta + 1\} \tag{4}$$

where T is the video length. Similar to the encoder, the decoder consists of the self-attention layer $Q = K = Z_i + E_f\ , V = Z_i, \widehat{Z}_\iota = Att_{dec.}(Q, K, V\ )$, where E_f is the frame encoding.

4. Proposed method

Based on the dataset obtained as described above, this paper proposes a framework for human activity recognition. Initially, the video is segmented into individual frames. In the second step, a scene graph generation algorithm is utilized to generate graph data, which is then converted into a graph data structure for storage. The third step involves re-embedding the generated data using a graph attention neural network. Subsequently, position encoding is applied to the new embeddings of each frame. Finally, an attention mechanism is employed to capture the relationships within the data, ultimately facilitating the recognition of human activities. Figure 1 shows the architecture of the human activity recognition framework.

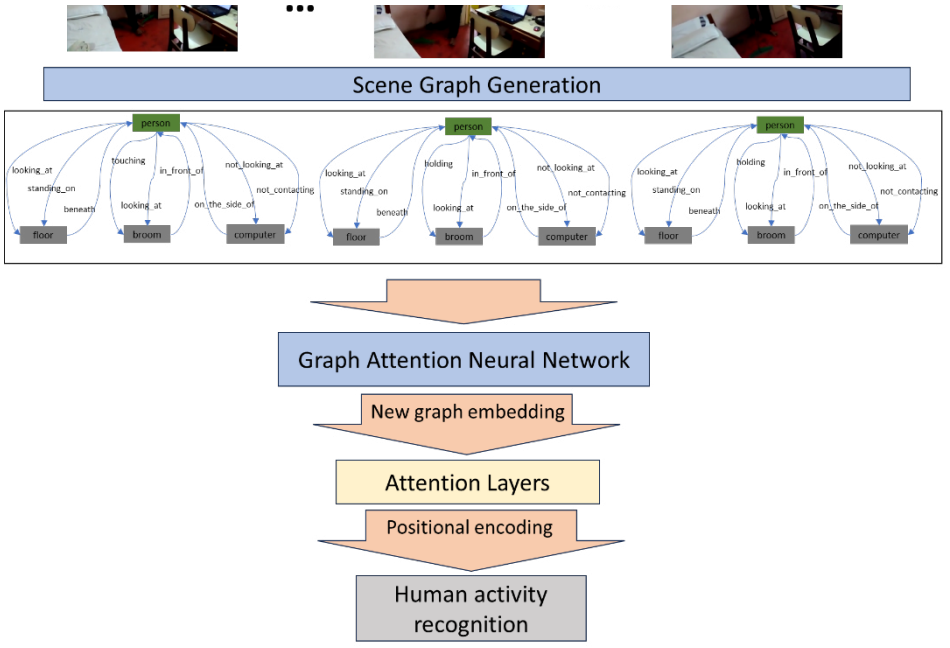


Fig.1. Human activity recognition architecture.

After obtaining the graph data, the graph attention neural network is capable of learning about the nodes and edges, subsequently re-embedding them. The computational formulas for the nodes and edges are as follows.

$$\alpha_{ij} = \frac{exp\left(LeakyReLU\left(\vec{a}^T[W\vec{h}_i||W\vec{h}_j]\right)\right)}{\sum_{k\in N_i} exp\left(LeakyReLU\left(\vec{a}^T[W\vec{h}_i||W\vec{h}_k]\right)\right)} \tag{5}$$

$$e_{ij} = \frac{exp\left(LeakyReLU\left(\vec{b}^T[W\vec{e}_i||W\vec{e}_j]\right)\right)}{\sum_{k\in N_i} exp\left(LeakyReLU\left(\vec{b}^T[W\vec{e}_i||W\vec{e}_k]\right)\right)} \tag{6}$$

where W is the training weight. || is used to denote a concatenation of two vectors. $\vec{a}$ and $\vec{b}$ is used for transforming the size of vector. The attention coefficients are then normalized using a softmax function to ensure that they sum to one. α_{ij} and e_{ij} are the new nodes and edges after computing.

Since our graph data is a sequence of coherent frames, we need to connect them. As each frame is represented by a different vector in the spatial part, we use position encoding to encode each frame so that the algorithm can learn the temporal features.

The self-attention mechanism operates by performing a scaled dot-product on the queries, keys, and values of all nodes. It then computes attention scores for each position by applying the SoftMax function to the compiled data. The attention is determined in the following manner:

$$A(Q,K,V) = Softmax\left(\frac{\bar{Q}K^T}{\sqrt{d}}\right)V \tag{7}$$

Where the $Q \in \mathbb{R}^{l_Q \times d_{in}}$, $K \in \mathbb{R}^{l_K \times d_{in}}$, and d_{in} is the input dimension. q_i, k_i, v_i represent the i_{th} row in Q, K, V, then the i_{th} query attention is:

$$A(q_i,K,V) = \sum_j \frac{k(q_i,k_i)}{\sum_l k(q_i,k_i)} v_j = \sum_j p(k_j|q_i)v_j = E_{p(k_j|q_i)}[v_j] \tag{8}$$

Each frame has a q and k value, which is derived from the attention mechanism. Although dot multiplication is required for each frame, in reality, not every frame is important, so important frames need to be selected for dot multiplication to improve prediction accuracy and calculation efficiency. Query's attention towards other keys is notated as a probability $p(k_j|q_i)$. If the $p(k_j|q_i)$ more like the $\frac{1}{l_K}$, it should be the element which can be ignored.

$$KL(q \parallel p) = \ln \sum_{j=1}^{l_K} e^{\frac{q_i k_l^T}{\sqrt{d_{in}}}} - \frac{1}{l_K} \sum_{j=1}^{l_K} \frac{q_i K_j^T}{\sqrt{d_{in}}} - lnl_K \tag{9}$$

Dropping the constant, the i-th query's sparsity measurement is:

$$M(q_i,K) = \ln \sum_{j=1}^{l_K} e^{\frac{q_i k_l^T}{\sqrt{d_{in}}}} - \frac{1}{l_K} \sum_{j=1}^{l_K} \frac{q_i K_j^T}{\sqrt{d_{in}}} \tag{10}$$

The i-th query is the $M(q_i,K)$, it should be the more important frame.

We have the ProbSparse self-attention by allowing each key to only attend to the dominant queries:

$$A(Q,K,V) = Softmax\left(\frac{\bar{Q}K^T}{\sqrt{d}}\right)V \tag{11}$$

where $\bar{Q}$ is a sparse matrix that matches the size of Q and it only includes the top-u queries as determined by the sparsity measure $M(q,K)$.

$$X_{j+1}^t = MaxPool(ELU(Conv1d([x_j^t]_{AB}))) \tag{12}$$

where X_{j+1}^t is the output of the multi-headed ProbSparse self-attention layer in this layer; $[x_j^t]_{AB}$ is the calculation result of the multi-headed ProbSparse self-attention layer in the previous layer; ELU is the activation function.

For fast prediction, we put the input X_{de} into the decoder layer.

$$X_{de}^t = Concat(X_{token}^t, X_0^t) \tag{13}$$

For the decoder layer, where X_0 is a placeholder of the target to fill the values with zeros, X_{token} is the start input which is same as the encoder layer. A fully connected layer acquires the final output. Till now, the kernel of temporal part has been created.

5. Result

We divide the data into 80% and 20% of samples for respectively the training and the testing sets. The loss function used is Cross-Entropy. The training and testing accuracies are shown in Figure 2. After 200 epochs of training, the accuracy can reach 0.86.

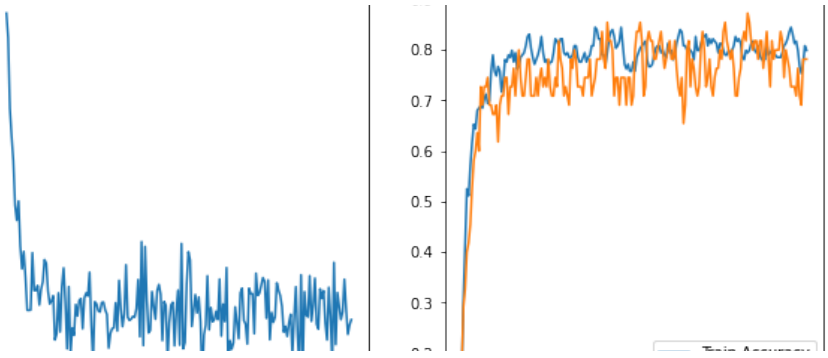


Fig.2. Human activity recognition accuracy.

Figure 3 presents the confusion matrix for the test set. We can see that for the test data, most of the classifications are accurate. But it must be mentioned that there are some reasons why some individual classifications are inaccurate.

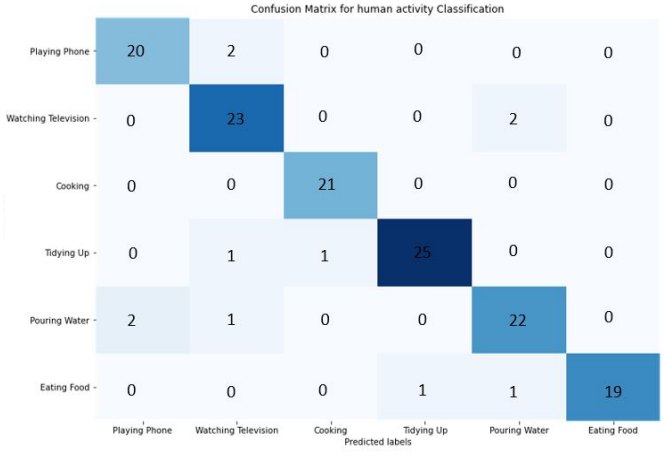


Fig.3. Human activity recognition confusion matrix.

Eating food can be confused with pouring water and tidying up because the same objects and relationships between human and objects often appear in these activities. For example, in pouring water, the cup is an important feature, but in eating food, the cup is also an object that often appears with food, so the two can be confused to a certain extent. However, the relationship between people and objects is not exactly the same, especially in videos. During eating food activities, although people may glance at the cup or even touch the cup, most of the time it is interactions between human and food. The key relationships between people and objects during the whole duration of the videos is important for identifying the activities correctly. Therefore, the context-based human activity recognition proposed in

this study is effective in eliminating ambiguity. Especially for similar scenes with the same object features, the contextual relationship is that special feature.

The analysis clearly demonstrates that the cooking activity is correctly identified in all videos. This accuracy is primarily due to the activity being closely associated with the stove, which is the central object used in cooking. This focus on the stove aids in disambiguating the activity, despite the presence of other objects like cups in the kitchen. In scenarios where the human is looking at the stove, it is a significant indicator of cooking activity.

To demonstrate the superior performance of our algorithm, we conducted comparative analyses with two leading algorithms currently used for spatial-temporal graph data. The results, presented in Table 2, clearly show that our algorithm outperforms the others.

Table 2. Comparison of results.

Algorithms	Loss value	Accuracy
Spatial temporal GAT	1.64	0.76
Spatial temporal GCN	2.1	0.80
Our algorithm	1.7	0.86

This model focuses on extracting temporal and spatial features. In addition, it learns the contextual relationship between people and objects, along with the characteristics of dynamic images in consecutive frames of the video. Finally, the attention mechanism in this model can better re-embed the features to learn their importance, thereby achieving higher accuracy.

6. Conclusion

This paper primarily investigates human activity recognition based on contextual relationships. Although there are numerous methods for recognizing human activities, such as those based on sensors and computer vision, these often overlook the significance of context. Currently, no publicly available datasets adequately store the relationships between people and surrounding objects. Consequently, this study initially selected videos with clearly labelled human activities from public sources to generate graph-structured data, which effectively determines the relationships between people and objects. This has led to the creation of a new graph dataset for human activity recognition. Furthermore, this research introduces a new framework architecture for human activity recognition. The results indicate that recognizing human activities, with an integration of context, achieves high accuracy and also performs well in disambiguation.

Effective reasoning and Interpretability can be achieved based on graph structures and this model. For future development, there is potential to collect more data to improve performance across a broader range of activity classifications. As the amount of data increases, including more types of human activities will make the model more generalizable. The real-time monitoring and further need for disambiguation that comes with that may be the next challenge. In addition, for outdoor human activities, there may be more complex context relationships and objects, which is also a challenge for future research.

Acknowledgements

The authors would like to acknowledge the support of the Centre for Artificial Intelligence, Robotics and Human-Machine Systems (IROHMS), operation C82092, part-funded by the European Regional Development Fund (ERDF) through the Welsh Government.

References

[1] Filippini Massimo, and Lester C. Hunt. (2011) “Energy demand and energy efficiency in the OECD countries: a stochastic demand frontier approach.” *Energy Journal* **32** (2): 59–80.

[2] Poppe Ronald. (2010) "A survey on vision-based human action recognition." *Image and vision computing* **28**(6): 976-990.

[3] Majd Mahshid, and Safabakhsh Reza. (2020) "Correlational convolutional LSTM for human action recognition." *Neurocomputing* **396**: 224-229.

[4] Shibo Zhang,Yaxuan Li,Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa. (2022) "Deep learning in human activity recognition with wearable sensors: A review on advances." *Sensors* **22**(4): 1476.

[5] Nicholas D. Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T. Campbell. (2010) "A survey of mobile phone sensing." *IEEE Communications magazine* **48** (9): 140-150.

[6] Wang, Zhihua, Zhaochu Yang, and Tao Dong. (2017) "A review of wearable technologies for elderly care that can accurately track indoor position, recognize physical activities and monitor vital signs in real time." *Sensors* **17** (2): 341.

[7] Claudio Bettini, Gabriele Civitarese, and Riccardo Presotto. (2020.) "Caviar: Context-driven active and incremental activity recognition." *Knowledge-Based Systems* **196**: 105816.

[8] Fuqiang Gu, PictureMu-Huan Chung, PictureMark Chignell, PictureShahrokh Valaee, PictureBaoding Zhou, and PictureXue Liu. (2021). "A survey on deep learning for human activity recognition." *ACM Computing Surveys (CSUR)*, **54**(8): 1-34.

[9] Tong Tong, Rossitza Setchi, and Y. Hicks. (2022) "Context change and triggers for human intention recognition." *Procedia Comput* **207**: 3826-3835.

[10] Ji Shuiwang, Xu Wei, Yang Ming, and Yu Kai. (2012) "3D convolutional neural networks for human action recognition." *IEEE transactions on pattern analysis and machine intelligence* **35**(1): 221-231.

[11] Wang Pichao, Li Zhaoyang, Hou Yonghong, and Li Wanqing. (2016) "Action recognition based on joint trajectory maps using convolutional neural networks." *ACM international conference on Multimedia* pp: 102-106.

[12] Fullerton E, Heller B, and Munoz-Organero M. (2017). "Recognizing human activity in free-living using multiple body-worn accelerometers." *IEEE Sensors Journal* **17**(16): 5290-5297.

[13] Sharma V, Gupta M, Pandey A. K, Mishra D, and Kumar A. (2022). "A review of deep learning-based human activity recognition on benchmark video datasets." *Applied Artificial Intelligence* **36**(1)" 2093705.

[14] Wang LuKun, and RuYue Liu. (2020) "Human activity recognition based on wearable sensor using hierarchical deep LSTM networks." *Circuits, Systems, and Signal Processing* **39** (2)" 837-856.

[15] Kaixuan Chen, Lina Yao, Dalin Zhang, Xianzhi Wang, Xiaojun Chang, and Feiping Nie. (2019). "A semisupervised recurrent convolutional attention model for human activity recognition." *IEEE transactions on neural networks and learning systems* **31**(5): 1747-1756.

[16] Bin Ren1, Mengyuan Liu, Runwei Ding, and Hong Liu. (2020). "A survey on 3d skeleton-based action recognition using learning method." *Cyborg and Bionic Systems*.

[17] Xiaojun Chang, Pengzhen Ren, Pengfei Xu, Zhihui Li, Xiaojiang Chen, and Alex Hauptmann. (2021). "A comprehensive survey of scene graphs: Generation and application." *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(1): 1-26.

[18] Fabian Caba Heilbron, Wayner Barrios, Victor Escorcia, and Bernard Ghanem. (2017). "Scc: Semantic context cascade for efficient action detection." *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3175-3184.

[19] Tao He, Lianli Gao, Jingkuan Song, Yuan-Fang Li. (2021). "Exploiting scene graphs for human-object interaction detection." *IEEE/CVF International Conference on Computer Vision* 15984-15993.

[20] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, Devi Parikh. (2018). "Graph r-cnn for scene graph generation." *European conference on computer vision (ECCV)* pp: 670-685.

[21] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, Felix Heide. (2021). "Neural scene graphs for dynamic scenes." *Conference on Computer Vision and Pattern Recognition* pp: 2856-2865.

[22] Sijie Yan, Yuanjun Xiong, and Dahua Lin. (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition." *AAAI conference on artificial intelligence* **32**, (1).

[23] Negar Heidari, Alexandros Iosifidis. (2021). "On the spatial attention in spatio-temporal graph convolutional networks for skeleton-based human action recognition." *International Joint Conference on Neural Networks (IJCNN)* pp: 1-7.

[24] Jinrui Yang, Wei-Shi Zheng, Qize Yang, Ying-Cong Chen, and Qi Tian. (2020). "Spatial-temporal graph convolutional network for video-based person re-identification." *IEEE/CVF conference on computer vision and pattern recognition* pp: 3289-3299.

[25] Chaopeng Pan, Haotian Cao, Weiwei Zhang, Xiaolin Song, and Mingjun Li. (2021). "Driver activity recognition using spatial - temporal graph convolutional LSTM networks with attention mechanism." *IET Intelligent Transport Systems* **15**(2): 297-307.

[26] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. (2020). "Action genome: Actions as compositions of spatio-temporal scene graphs." *The IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp: 10236-10247.

[27] Yuren Cong, Wentong Liao, Hanno Ackermann, Bodo Rosenhahn, and Michael Ying Yang. (2021). "Spatial-temporal transformer for dynamic scene graph generation." *The IEEE/CVF international conference on computer vision* pp: 16372-16382.