International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Deep Learning-Based Group Activity Recognition Using Multiperson Relational Graph

Smita S. Kulkarni*[1], Sangeeta Jadhav[2], Avinash N Bhute[3], Harsha A Bhute[4], Ashitosh D. Chavan[5]

Submitted:11/03/2024 **Revised**: 26/04/2024 **Accepted**: 03/05/2024

Abstract: The recognition of group activities (GAR) is of significant importance in the field of computer vision as it facilitates the investigation and understanding of patterns of human behavior. Existing methodologies mostly concentrate on interactions at the interpersonal level within a group. However, sociological research has emphasized the significance of individual characteristics, interactions at the multi-person level, and the overall structure of the group in recognizing group activities. Hence, in this research, to represent the relationships between people's locations and appearances, adaptable and effective multi-person relational graphs (MRG) have been developed for the aim of GAR. Graph Convolution Network (GCN) with sparse temporal sampling is applied to efficiently infer multi-person relational graphs. The proposed network distinguishes group activity from individual interaction via relational reasoning. The use of a GCN for identifying group activities comes after the implementation of a deformable CNN to collect features and categorize individual actions. For multi-level interaction reasoning and group structure modeling, visualization samples and experimental results show that this approach works better than the best methods currently available. These findings highlight the necessity of taking into account multi-person relational graphs (MRG) representations for recognizing group activities.

Keywords: *Group Activity Recognition, Graph Convolution Neural Networks, Deformable CNN*

1. Introduction

The multimedia community has shown significant interest in GAR [1] due to its wide range of practical applications in areas such as intelligent surveillance, sports video understanding, and social behaviour analysis. Several previous studies have shown that it takes a lot of attention to tell the difference between actions done by a single person or complicated human activities in a video [2]. Understanding social interactions and behaviour in real-world scenarios requires the ability to identify group activities. In the field of computer vision, identifying group activities is a challenging task since the objective is to evaluate and understand the collective behaviours of multiple people [3] [4] [5] [6] [7]. To identify activities that involve several individuals, it is essential to capture both the interactions between persons [8] and the spatiotemporal information at the individual level. Recognition is classified into two types: activity recognition and group interaction recognition. Activity recognition identifies an individual's actions, whereas group interaction recognition identifies group interactions. Group activity recognition approaches integrate activity and interaction detection to better comprehend group behaviour. Figure 1 [4] depicts a group of people standing stationary within a frame. Observing only one person in the yellow bounding boxes (M) and (N) and disregarding their surroundings may reveal that they are simply standing about. To fully understand a group's activity, it's important to include all members of the group, including their positions and surroundings. To understand group dynamics, it's important to consider both the individuals and their settings.

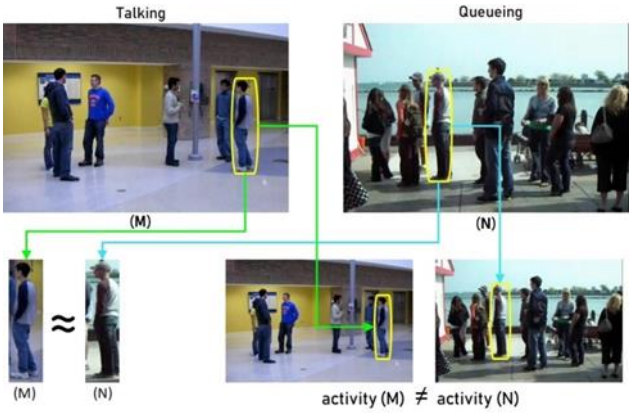

Fig. 1. Interaction among people in group activities: Talking and Queuing

In summary, the primary contributions of this study are as follows: 1) A deep learning model was proposed to improve the accuracy and computing efficiency of the GAR. This GAR framework takes into account the context of interactions between multiple people in order to capture both group and individual activity. The performance of individual action recognition is enhanced by proposing a Deformable Convolution MobileNet4 CNN. 2) To depict the relationships between people's locations and

[1, 5] *MIT Academy of Engineering, Pune, 412105, India.*
[1]*ORCID ID: 0000-3343-7165-777X* [5]*ORCID ID: 0000-0003-0962-4031*
[2] *Army Institute of Technology, Dighi, Pune, 411015*
ORCID ID: 0000-0002-0610-0374
[3, 4] *Pimpri Chinchwad College of Engineering, Nigdi, Pune, MH, India*
[3]*ORCID ID: 0000-0003-3236-686X* [4]*ORCID ID: 0000-0002-1522-4053*
** Corresponding Author Email: sskulkarni@mitaoe.ac.in*

appearances, adaptable and effective multi-person relational graphs (MRG) are constructed for the aim of GAR. It possesses the capacity to distinguish between distinct activities that include groups of individuals. 3) Effective actor relation graph inference is achieved through the use of GCN with sparse temporal sampling. The proposed framework uses relational reasoning to distinguish group activities from individual interactions.

2. Literature Survey

The process of identifying and classifying human activities that occur in a group setting is referred to as group activity recognition, such as in a meeting, sports game, or surveillance. This section offers a comprehensive review of the many strategies that are currently used to identify group activities. Recognizing group activities precisely can be challenging due to the complicated interactions and interdependencies between group members. Most human activity recognition reviews have focused on individual action recognition [9] [2] [10] [11]. In [12] presents group activity recognition in smart buildings, and [13] investigates various facets of human activity, although they only present conventional approaches. Group activity recognition was most recently surveyed in [14]. It emphasizes handcrafted methods, while [15] discusses deep learning approaches as well. Recently, a growing interest has been in developing intelligent systems that automatically recognize group activities. Recognizing group activities is difficult due to the complexity of group dynamics, the variety of human actions, and the environmental variability. Researchers have looked into both handcrafted traits and deep learning techniques to help with the recognition of group activities. With the increasing popularity of deep learning, many researchers have been using deep learning methods to recognize group activities. This section discussed various hand-crafted feature-based and deep learning-based group activity recognition techniques Traditional approaches [8], [16], and [17] often use custom features and set rules, which may not accurately reflect the complex temporal dynamics and interactions within the group. CNN-based Human Activity Detections using Profound Learning (HADPL) is a method presented in [18] this paper that detects HARs from acquired accelerometer data. The use of a transfer learning CNN model along with a RoI pooling layer in [19] this study shows how action recognition can be improved. Recent advancements in deep learning techniques have demonstrated encouraging results in the field of video-based group activity recognition. These methods typically use a two-stage recognition technique, as described in [20]. In the first stage, a convolutional neural network (CNN) is used to extract person-level information. Next, a global module has been created to aggregate these distinct images into a feature at the scene level. When the two-stage deep temporal model from [21] is used, an LSTM model is used to record patterns of individual activity while data is collected at the individual level. A unified framework for identifying and detecting numerous people's simultaneous activities is proposed in [22]. Using a hierarchical relational network that assigns a linked description to each individual was proposed [23]. Furthermore, some studies [6] [24] look at the usage of structured recurrent neural networks to represent the scene environment or generate captions [25]. A novel approach in deep learning research involves integrating graphical representations with deep neural networks. To address these constraints, researchers have utilized deep learning methods, specifically Graph Convolutional Networks (GCNs), [26], to enhance the accuracy of group activity recognition. However, to create a graph that is fully connected over the whole video frame, it would be impractical to compute all of the pairwise relationships that exist throughout all video frames. Construct a multi-person graph according to their relative locations. Moreover, for enhanced learning, suggest integrating GCN with a method of temporal sampling [27]. Develop a multiperson graph based on their relative locations. Furthermore, for improved learning, consider combining GCN with a temporal sampling method [27]. Based on other factors, such as the similarities in their looks or the positions they hold for one another, it is acceptable to assume that the individuals in question are related to one another. To recognize collective activity in a situation involving multiple people, it is necessary to model individual interactions. Because they can handle changing times, recurrent neural networks (RNNs) are often used to model how video frames change over time [21]. These models can be costly to compute and may not accurately reflect group activity variations. However, [28] solely analyzed actor interactions and did not consider explosive relationships between individuals when identifying group activities. In recent times, deep learning-based methodologies have exhibited encouraging outcomes in the domain of group activity recognition from videos [5], [29]. To begin, the Convolution Neural Network (CNN) is responsible for gathering information at the actor level. Using a recurrent neural network (RNN) or other aggregation module, the generalized autoregressive model (GAR) can make group-level representations. This is done by combining actor-level features with an RNN. The nodes provide a visual representation of each actor, while the edges display their interactions. An actor relation graph (ARG) that takes into account the connections between actors based on both their appearance and their location was created in response to this idea [28]. The attributes of each node cannot be dynamically updated via graph reasoning, though, because the matrix of relations in an ARG is fixed. ARG does not incorporate motion contexts from other modalities, such as optical-flow images, while creating graphs; instead, it exclusively employs visual cues

from raw RGB frames. Video group activities naturally incorporate a variety of visual contexts, including motion patterns and appearance signals. [30] Using a self-attention technique known as a transformer, RGB, optical flow, and posture presentations were altered to identify group activities [31]. This method essentially uses a late fusion mechanism to build prediction scores from multiple modalities without recording actor-level associations. Prior group-activity identification algorithms have been demonstrated to improve visual tasks, but they only make use of local visual context and fail to use global embeddings that rely on conceptual labeling relationships. To recognize group activities, this study investigates the problem of recording the relationship between people's locations and appearances. The major objective is to create a more flexible and effective model of individual interactions, which will enable efficient inference for group activity recognition as well as automatic learning of the graphical connections between actors from video data. The primary objective is to create an individual relationship model that is more flexible and effective, enabling the efficient execution of inference as well as the automatic learning of the graphical connections between actors using video data. A deep learning model is suggested as a backbone network to increase processing speed. This deep learning architecture consists of two stages: Graph Convolutional Networks (GCN) are used in stage two to distinguish group activity from multi-person relational graphs (MRGs), while stage one uses CNN to extract person attributes from video frames of collective activities. Utilizing a GCN to identify group activities comes after using a deformable CNN (Cheng 2020) to collect features and categorize individual actions. The suggested MRG uses representational modeling on the combined visual graph to get context-aware information about each person's attributes and figure out how they normally interact with others. Following this, a multi-user visual context module is constructed utilizing two-stage aggregation techniques to dynamically modify node properties. They acquire meaningful representations by propagating information across nodes, utilizing the graph topology. GCNs use graph convolutions to capture both local and global dependencies in the graph. GCNs effectively model group relationships and interdependence when recognizing group activities.

3. Action Recognition Framework

The approach includes a pre-trained CNN model with transfer learning. Pre-trained layers extract visual characteristics and allow for weight updates to improve action classification. The transfer learning approach as shown in Figure 2 demonstrates the use of a pre-trained CNN architecture for feature vector extraction. In video sequences with numerous people, the CNN deep model is used to recognize individual actions. This research uses pre-trained transfer learning models with modified layers, such as VGG16, InceptionV3, and MobileNet. MobileNet outperforms other prominent CNN models in recognizing individual actions. The MobileNet model optimizes for individual action aspects, resulting in great efficiency and accuracy.

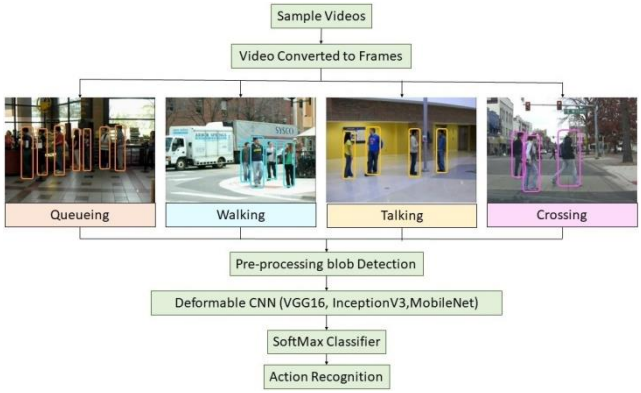

Fig. 2. Action recognition framework

3.1. DCNN-Based Feature Extraction

Individual action recognition is not very accurate since [100] only uses the region of interest and CNN to achieve action identification; instead, it largely focuses on group activity detection. As shown in Figure 5.1, this study suggests using a deformable convolution neural network (DCNN) to fine-tune the main network and make predictions for both individual and group activities more accurate. A key process in deep learning, especially in CNNs, is convolution. It entails applying filters or kernels to modify an input image or feature map. CNNs use convolutional layers to apply filters that extract important features from the input. Detecting human actions in videos presents numerous difficulties. The temporal and spatial variations that take place inside actions present a significant problem. The human body may deform in these variants in ways including bending, stretching, or occlusions brought on by other things. The insignificant capture of these deformations by conventional convolutional layers results in inadequate action recognition ability. However, they are computationally expensive when regions are multiplied deeper into the network. Deep convolutional neural networks (DCNN) can get around the problems that regular convolutional layers have by adding spatial deformations to the convolutional process. Instead of assuming a fixed receptive field, deformable convolution lets the network learn the sets for sampling the input characteristics in a way that is best for them. The sampling locations of the convolutional filters are set by offsets, which are extra parameters that can be learned in deformable convolution. These offsets are determined depending on the input features during the training stage. Better adaptation to spatial deformations in

human actions is possible with deformable convolution.

3.2. Implementing Deformable Convolution

Deformable convolution can be used to recognize human actions by adding layers to existing CNN structures. These layers can be substituted for or used in addition to regular convolutional layers in the network. The model can be trained to make use of the advantages of deformable convolution by fine-tuning the network on action recognition datasets. The proposed approach makes use of DCNN to solve the previously mentioned issue. As shown in the formula that follows, this makes it possible to combine information between a specific pixel position and its surrounding pixels.

$$y(i,j) = \sum_{i,j} \omega_{i,j} P(x_i, y_j) \quad (1)$$

For each frame, the sample expression for the *(i, j)* position, where *y(i; j)* stands for DCNN, $\omega_{i,j}$ describes the weight, and $P(x_i, y_i)$ is the value of the pixel. The model selects diverse features by obtaining information about offset pixel position and creating a deformable feature map structure. The accuracy increases with the amount of information added to the classification because more pixel-level data collected during the pooling process is combined. Deformable convolution offers various advantages for human action recognition systems. First, it improves the ability to record fine-grained spatial information and local deformations, resulting in more discriminative action representation. Second, it strengthens action recognition models against occlusions and viewpoint alterations. Finally, it improves generalization by accommodating different action speeds and scales.

4. Recognizing Group Activities Using The GCN Model

It is necessary to model interpersonal interaction to identify collective behaviour in a multi-person setting. The goal of this approach is to effectively learn about the discriminative relationship between people using deep models. Inferences on MRG can be easily carried out using regular matrix operations, and MRG connections can be automatically learned by GCN through group activities. Furthermore, for efficient video modelling, localized and temporally randomized MRG provides effective sparsification techniques.

In Figure 3, the CNN model is proposed for individual actions, followed by a graphical convolution neural network (GCN) [32] for group activity recognition. The Deformable Convolution module is recommended in this work to improve CNN's transformation capabilities. This module can effectively learn several geometric transformations from video frames.

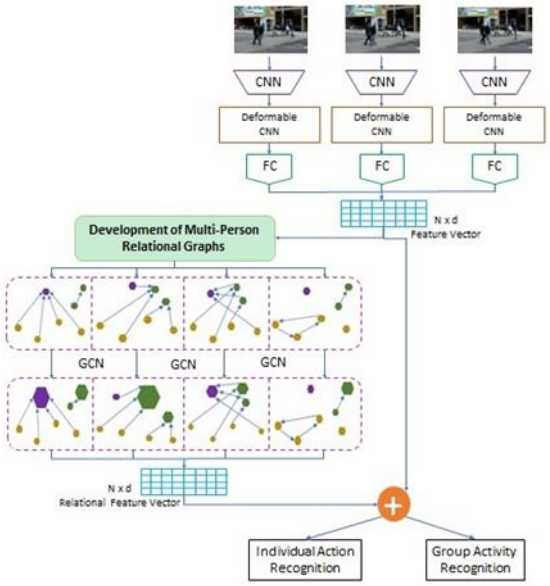

Fig. 3. Framework of group activity recognition using GCN

The foundation of this method is adding offsets to the spatial sample locations. Back-propagation combined with deformable convolutional networks makes it robust to train with the backbone CNN model. The major goal is to recognize group activity in a scene using relational information [33], [32], and [5]. To accomplish this objective, developed a multi-person relational Graph (MRG) that represents the interaction between several individuals. Subsequently, relational reasoning on this graph through GCN model is implemented to recognize group activity. The system framework is illustrated in Figure 5.2. This framework's deep learning model consists of two stages. The first stage uses CNN as its backbone, the model extracts personal feature vectors from video frames in stage one. Subsequently, GCN is employed in stage two to identify group activity from multi-person relational graphs. The DCNN technique used to extract the features for each person's bounding boxes from the frame feature map. To obtain an n-dimensional vector of features for every individual, an FC layer is applied to the matching features. The total number of bounding boxes in V frames is denoted by M. An individual's feature vectors are represented by the M*n matrix as R. Next, build multiperson relational graphs with each node representing an actor based on their unique qualities. Each edge in the graph reflects a scalar weight determined depending on the multiperson aspect and position. Multiple relation graphs are created from the same set of features to represent different relationship information. Individual and group activities are recognized through learning and inference. GCN is used to calculate relational context based on MRG. After that, the graph convolution and MRGs are merged to create the M*d dimension matrix R, which represents individual relationships. To improve individual action classification, add a fully connected layer to each person's

representation. Max-pooling multiperson representations results in scene-level representations. A fully connected layer above it detects group activity.

5. Development of Multi-Person Relational Graphs

The graph structure is employed to represent multi-person relational information. The relational reasoning applied to GCN was inspired by [28] [33] to classify group activity. In the graph structure individual is represented as node $I = (u_x^a, u_x^c | x = 1, \dots \dots P)$, where P is the number of people,

$u_x^a \varepsilon \mathbb{R}^d$ is individual x's aspect feature and center coordinate of individual x's within bounding box is $u_x^c = \left(b_x^i, b_x^j\right)$. The graph $G\varepsilon\mathbb{R}^{PxP}$construe to represent multi-person relational context among people and $G_{x,y}$ indicates a relational graph. To represent the relationship between multiple people aspect features and position information must be taken into consideration. Furthermore, the conceptual characteristics of the aspect and position relations are distinct. To that purpose, need to describe the appearance and position relations explicitly. The following combination function defines the relational context value in (2).

$$G_{x,y} = k\left(f_a\left(u_x^a, u_y^a\right), f_c\left(u_x^c, u_y^c\right)\right) \quad (2)$$

where $f_a\left(u_x^a, u_y^a\right)$ indicated the aspect relation among people and position relation is estimated by $f_c\left(u_x^c, u_y^c\right)$. The function k combines the appearance and position relational context feature into a scalar weight. In the experimentation, the following function in (3) is used to estimate the relational context.

$$G_{x,y} = \frac{f_s(u_x^s, u_y^s) exp\left(f_a(u_x^a, u_y^a)\right)}{\sum_{y=1}^{P} f_c(u_x^c, u_y^c) exp\left(f_a(u_x^a, u_y^a)\right)} \quad (3)$$

where in, to ensure optimal results, it is necessary to normalize the input data before applying the SoftMax classifier. Normalization is the process of transforming data to a standard scale, eliminating variations, and bringing all features or variables onto a similar range. It is particularly important when dealing with datasets containing attributes with different scales, units, or distributions. Normalizing the data, makes it easier for learning algorithms to process and compare the features accurately.

5.1. Aspect Relation

Aspects of human appearance are crucial for recognizing group activities. These attributes offer significant insights into the identities, activities, and interactions of individuals within a group setting. Visual signals from the human body, such as shape and posture, are vital for comprehending group dynamics. Pose estimation algorithms can recognize and track body joints, resulting in skeletal structure representations. Skeletal structure representations play a crucial role in skeleton-based action recognition, a process that deduces the activities performed from the movements and gestures of individuals. This method is effective even in situations involving challenging illumination conditions or partially occluded individual appearances.

1. Dot-Product

The mathematical process known as the dot-product is used to determine the degree of similarity that exists between two vectors. Measurement of the degree of similarity between various activity patterns and the identification of commonalities among them are two of the most important functions that it performs in the context of GAR. Equation (4), which uses the dot-product to assess the degree of similarity across activity patterns, enables us to differentiate between various activities and recognize group behaviours

$$f_a\left(u_x^a, u_y^a\right) = \frac{(u_x^a)^T (u_y^a)}{\sqrt{\alpha}} \quad (4)$$

where α is the normalization factor.

2. Embedded Dot-Product

In the context of group activity recognition, vectors represent activity patterns. These vectors are embedded into a high-dimensional space using techniques such as deep learning models or dimensionality reduction algorithms as described in the equation. By computing the dot-product [34] between embedded representations, can assess their similarity and determine whether they belong to the same group activity or not.

$$f_a\left(u_x^a, u_y^a\right) = \frac{\psi(u_x^a)^T \psi(u_y^a)}{\sqrt{\alpha_l}} \quad (5)$$

Where $\psi(u_x^a) = \omega_\psi u_y^a + b_\psi$

$\omega_\psi \varepsilon \mathbb{R}_l^d$ is weight matrices and $b_\psi \varepsilon \mathbb{R}_l^d$ is bias vector.

3. Relational Graph

The relational graph is evaluated as given in equation (5).

$$f_a\left(u_x^a : u_y^a\right) = ReLU\left(\omega\left[\psi(u_x^a), \psi\left(u_y^a\right)\right] + b\right) \quad (6)$$

In equation (6) ω and b are learnable parameters and ReLu activation function introduces nonlinearity to deal with the complicated representation.

5.2. Positional Relation

An essential component of action recognition involves comprehending the positional relationship among individuals. The concept of position relation relates to the spatial configuration and comparative distances among individuals. It imparts significant insights into the intricacies of motion dynamics and patterns of interaction throughout an action. By capturing and encoding the

position relationship with precision, it becomes feasible to distinguish between various actions and enhance the overall performance of recognition. Identifying the relative positions and movements of body components is essential for action recognition algorithms. In human action recognition, the Euclidean distance is a commonly used metric to characterize the spatial relationships between body components. Action recognition algorithms are capable of acquiring the spatial information required by computing the Euclidean distance between pairs of body parts.

As per the observation, $G_{x,y}$ set as per the threshold value in (7)

$$f_s(u_x^s : u_y^s) = \prod(d(u_x^s : u_y^s) \leq \lambda) \quad (7)$$

Where π is the indicator performance and $d(u_x^s : u_y^s)$.

Action recognition algorithms can discriminate between action classes by evaluating spatial interactions and extracting discriminative features. After localizing individuals, pairwise Euclidean distances are obtained between pairs. A hyper-parameter l serves as a distance threshold and specifies the distance between the centers of two distinct bounding boxes in Euclidean terms. The encoding of Euclidean distance remains robust to changes in orientation and posture. It enables the body part arrangement to be the primary focus of the recognition system instead of the absolute positions of the body parts. The encoding of Euclidean distance remains robust to changes in orientation and posture. It enables the body part arrangement to be the primary focus of the recognition system instead of the absolute positions of the body parts.

5.3. Multiple Relational Graphs

Based on various relationships of interest, many relational graphs are built. For instance, one graph may show spatial interactions between individuals, while another shows temporal interdependence between actions. Hierarchical group structures and social interactions can be represented by additional graphs. Building a multi-person relational graph (MRG) to store different kinds of relationship information is important when looking at context information between multiple people. The development of several graphs on individuals is represented by the equation $G = (G^1, G^2, \ldots\ldots, G^{Pg})$ where P_g is the number of graphs. By fusing or combining the multiple relational graphs, a unified representation of the group dynamics is produced. The information from various graphs may be merged via concatenation, averaging, or more complex aggregation methods during this fusion procedure. A relational graph is a visual representation that contains distinct relationships or interactions, providing significant contextual information. The integration of various graphics provides an improved representation of the group context, facilitating enhanced differentiation among distinct activities.

5.4. Temporal Modeling

One of the most significant challenges in the field of group activity recognition is the identification of the temporal dynamics and interactions that occur between individuals across time. This article examines the importance of temporal modelling in GAR and how multiple graphs might improve recognition accuracy and robustness. Capturing the temporal dynamics of events and interactions over time is known as temporal modelling. Insight into the temporal contexts, dependencies, and sequential patterns necessary for precise activity detection is made possible. Recurrent Neural Networks (RNNs) are extensively employed in the field of sequential data modelling due to their ability to effectively capture temporal dependencies among different graphs. This enables the network to identify long-term dependencies and make predictions based on the whole temporal context. Sparse sampling is the process of randomly selecting frames from the video to create the temporal graph. This strategy reduces computation time and computation costs. The proposed GCNs are deep learning models that function with graph-structured data. They collect local and global dependencies and aggregate information from surrounding nodes using graph convolution algorithms. Adding temporal information to GCNs made it easier to come up with ways to model dynamic graphs and use temporal reasoning in tasks like activity recognition

6. Graph Inference and Learning

In GAR, a collection of individuals is depicted as a graph, with each individual serving as a node and the relationships between them as edges. The mathematical representation of a graph is given by the equation $G = (V; E)$, where V denotes the set of nodes (individuals) and E denotes the set of edges (relationships). Graph convolution aggregates information from surrounding nodes in a graph. Graph convolution can be explained mathematically using (8):

$$H^{(l+1)} = \sigma(G\ H^l\ W^l) \quad (8)$$

This multiplication represents the flow of information between nodes based on their relationships. Where $G\varepsilon\mathbb{R}^{(NXN)}$ is the graph's matrix representation. $H^l\varepsilon\mathbb{R}^{(NXd)}$ is a representation of a node's features in the l^{th} layer and $W^l\varepsilon\mathbb{R}^{(dXd)}$ is the weight matrix, which needs to be learned as layerspecific. The weight matrix W comprises learnable parameters that establish the significance of neighbouring node information for every node in the network. To improve model performance, these parameters are adjusted throughout training. A weight matrix enables the model to learn and adjust the most important features for accurate group activity recognition.

The activation function is responsible for introducing non-linearity into the graph convolution procedure, which enables the model to acquire knowledge of intricate patterns and complicated relationships. The activation function σ selected for the group activity recognition assignment is determined by its particular characteristics and requirements. In real life, several graph convolution layers are stacked on top of each other to record information about hierarchies and help the model learn more general representations. Iteratively, each layer does the graph convolution process, and the output of one layer is fed into the input of the next layer. The iterative process employed in this study enables the model to gradually enhance its comprehension of group activities by incorporating insights from both local and global contexts. Following a number of graph convolutional layers, pooling techniques can be used to collect data from various nodes and identify the overall characteristics of the group activity. Mean pooling and max pooling are two popular pooling algorithms that combine the properties of the nodes into a fixed length representation. After the features are combined, the data is sent to a classifier, which could be a fully connected neural network or a softmax layer, so that the group's activities can be put into groups. During the classification process, the classifier either learns to map the pooled features to the activity labels that correspond to them or predicts the activity probabilities for various classes. Consider the concatenation operation as a fusion function as well in (9).

$$H^{(l+1)} = \sum_{i=1}^{N_g} \sigma\left(G_i\, H^l\, W^{(l,i)}\right) \qquad (9)$$

Additionally, examine the concatenation operation as a fusion function. Alternately, a process known as early fusion can combine a collection of graphs into a single graph. This implies that they can be summed up ahead of the GCN process to combine them into a single graph. In the concluding step, the output relation features from GCN are aggregated with the initial features to produce the scene illustration. As shown in Figure 3, the representation of the scene is given as input to two different classifiers, which then yield individual action and group activity predictions, respectively.

6.1. Inference and Learning

The model learns to predict the activity label based on the node properties and graph topology. Several techniques, including gradient descent and backpropagation, are applied during the training phase to maximize the model's parameter settings. The trained GCN model predicts the label for a group activity based on its input graph. The model successfully recognizes group activity based on temporal dependencies and interactions. Cross-entropy loss is a commonly used loss function in classification applications, such as group activity recognition. It measures the difference in predicted activity probabilities. The model is driven to minimize the discrepancy between the actual and predicted probabilities by the cross-entropy loss. Backpropagation gradients are used to optimize GCN model parameters, minimizing crossentropy loss. Iteratively changing parameters based on gradients improves the model's effectiveness in recognizing group activities. The final loss function is generated by combining the standard cross-entropy loss with it in (10).

$$\ell = \ell_1(y^G, y^G) + \rho\ell_2(y^l, y^l) \qquad (10)$$

In equation (10) ℓ_1 and ℓ_2 are cross-entropy loss functions, and ρ is utilised to optimize the training task over y^G group activity and y^l individual action.

7. Experimental Results

Extensive experimentation was undertaken to validate the effectiveness of the proposed methodology. The datasets and implementation details utilized in the experiments are presented. In contrast to state-of-the-art (SOTA) methodologies, proposed experimental findings are compared to various interaction modules. In conclusion, we analyze and present visualizations of our findings to intuitively demonstrate their efficacy.

7.1. Experimental Data Sets

The investigations utilize two distinct data sets: the Collective Activity dataset (CAD) [4], and the Collective Activity Extended dataset (CAED) [35]. In addition to the pre-existing identifiers, we annotated the top-view positions and orientations of the individuals in preparation for MRG modelling. CAD: CAD is a collection of 44 videos organized into social scenarios. Each person has a bounding box and labels for their activity and stance. Action labels: NA, moving, waiting, queuing, and chatting. Pose labels indicate specific orientations, such as right, front-right, front-left, left, back-left, back, and back-right. Each ten-frame video clip is labelled as a group activity based on the majority of individual activities.

CAED: A CAD extension version is called CAED. It updates the videos with more labels for jogging and dancing. Pose labels are marked in the same way as CAD and correspond to specific orientations.

7.2. Implementation Details

The experiments were conducted using the PyTorch framework having CUDA cores on four NVIDIA Professional Series Quadro P6000 RTX PCIe 3.0 - 24GB GPUs. Hardware machine with 2X Intel Xeon silver 16 cores processor, 128GB. The network is trained with a dropout ratio of 0.3, a learning rate of 0.00001, and in 100 epochs with a mini-batch size of 32. In order to learn the network parameters while keeping the hyper-parameters fixed, make use of stochastic gradient descent using

ADAM. The GCN parameter is selected as dk = 256, ds = 32, and the distance mask threshold of 1/5. Based on each person's ground-truth bounding boxes, extract a 1024 feature vector using the methods outlined in Section 5.4. The default backbone CNN network, Inception v3, is utilized for feature extraction. Additionally, the experiments make use of CNN models that MobilNet maintains. Fine-tune the pre-trained ImageNet model on a single randomly chosen frame from each video, without using GCN. Subsequently, adjust the weights of the network's feature extraction section, and continue using GCN to train the network.

7.3. Comparison with State-of-the-Art (SOTA) Methods

The results of the experiments are shown in Table 1, 2, and 3. They show that our method achieves performance levels that are comparable to the best existing techniques (SOTA). In the subsequent section, we will explore the benefits of our technique by analysing its distinctive features in contrast to the state-of-the-art (SOTA) methodologies. A deep learning model is offered as a backbone network to enhance computing speed. This deep learning framework consists of two phases. In the first stage, the model utilizes Convolutional Neural Networks (CNN) to extract characteristics of individuals from video frames. In the second stage, Graph Convolutional Networks (GCN) are employed to identify and understand group activities based on the multiple-person relational graph. In Table 1, you can see a summary of the first set of activity recognition experiments and how well the CNN backbone model did in terms of computation time and accuracy. The MobileNet model is superior to the InceptionV3 model in terms of action recognition computation time. The efficacy of individual action recognition is enhanced through the implementation of DCNN transfer learning. When compared to models like Inception V3, MobileNet exhibits quantifiable performance improvements. Consequently, for group activity detection, the DCNN MobileNet is suggested as the foundational model for discriminating individual activities from each video frame. Since [100] is predominately concerned with the recognition of group activities, CNN and the region of interest are insufficient for accurate action recognition of individuals. The proposed DCNN model enhances the accuracy of individual action identification by fine-tuning the backbone network MobieNet and including skeleton extraction. Using deformable CNN allows for less data augmentation, which is a benefit. Deformable convolution allows the network to adaptively sample features from different locations, capturing spatial deformations in human actions. This enhances the network's ability to model complex motions and improves action recognition accuracy.

Table 1. Results of stage one for action recognition

Backbone Network	Overall Accuracy (%)	Time Complexity (%)
InceptionV3 [33]	90.91	78.67
MobileNet [33]	89.37	87.91
DCNN Inception V3	95.41	79.02
Proposed DCNN MobileNet	93.52	88.90

Conduct a comprehensive analysis of the CAD and CAE datasets. For the backbone networks, relationship modelling is suggested, and the assessment metric is the precision of group activity prediction. The results of the action recognition experimentation are presented in Table 2. Person action accuracy is used as the assessment metric in the proposed backbone network.

Table 2. Performance of CNN model for action recognition

Backbone Network	CAD Accuracy (%)	CAE Accuracy (%)
Multistream CNN [33]	83.15	-
VGG16	88.68	89.02
Inception V3	95.41	95.82
MobileNet	90.02	90.04
ROI Pooing MobileNet	91.79	92.08
Deformable MobileNet (proposed)	93.52	95.43

In the second step, a random selection of frames from each training example is used to fine-tune the backbone network that has already been trained. As the backbone network, MobileNet is used for the tests. In step two, the weights of the backbone network that is responsible for feature extraction are fixed. The GCN network trains to use integrated dotproduct to find the aspect and positional link as MRG. According to Table 3, the proposed backbone network DCNN MobileNet analyses the accuracy of group action prediction by utilizing person action as the evaluation measure.

A substantial amount of testing is carried out on two standard datasets for GAR. These datasets are known as the CAD dataset and the CAED dataset respectively. The performance of both datasets is considered to be state-of-the-art. The suggested MRG (Multi-Relational Graph) can effectively collect discriminative relational context information for group activity recognition.

Several techniques were used by these systems, namely RMIC [50], MLST-Former [37], P2CTMD [38], to reason about interactions for GAR. . RMIC modelled interactions at the person, group, and scene levels, while MLST-Former [37] prioritized inferences between bodily areas and individuals. Position information is utilized by P2CTMD [38] to improve the learning of pairwise relationships between individuals.

Table 3. Results of GCN model for GAR

Method	CAD Accuracy (%)	CAED Accuracy (%)
ARG CNN [28]	91.00	-
Improved ARG [33]	93.04	-
Fast Deep [36]	83.80	94.10
P2CTMD [36]	83.80	94.10
MLST-Former [37]	96.80	95.60
P2CTMD [38]	96.10	98.20
GCN(proposed)	97.40	98.20

However, these approaches failed to account for interactions at the subgroup level and neglected to capture the relationships of progressive reasoning across levels. As a result, the proposed strategy functions more effectively than conventional techniques by considering the relational graphs created by analysing a person's appearance and their positional relations in video frames. The GCN model is utilized to examine the multi-person relational graph-based framework. As a result, of this, it is possible to obtain higher levels of accuracy in recognizing individual and group activities in the proposed framework. The GAR framework combines individual and group actions by considering interaction context among several persons. Proposed Deformable Convolution enhances individual action recognition performance.

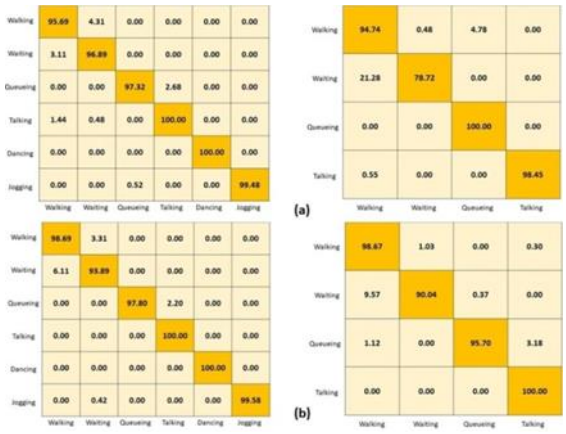

Fig. 4. P2CTMD and the proposed method are compared using the CAED (a) and CAD (b) datasets. The confusion matrix for P2CTMD is (a) whereas ours is (b).

Fig. 5. Visualization of interactions. Interactions are enhanced among persons exhibiting similar actions. Groups (a), (b), and (c) are engaged in the activities of moving, queuing, and dancing, respectively.

The exceptional performance of our method in identifying speaking, dancing, and jogging activities is illustrated in Figure 4 and compared with P2CTMD through the confusion matrix. Our method recognizes crossing and waiting actions more accurately and with less confusion. This superiority can be achieved through the effective utilization of the MRG modelling module, permitting a robust understanding of spatial patterns within groups. Specifically, in crossing groups, individuals are observed to be centrally distributed along the road, whereas in waiting groups, individuals are found to be distributed on one or both sides of the road.

8. Visualization

People from the same group usually act in similar ways. As shown in Figure 5, the pictures show person IDs and the actions that go with them. The contact weights between people are shown below the pictures. High interaction weights mean that these people have similar subgroup-level representations and tend to form groups that stick together. The fact that this happened shows that the interaction reasoning tool works to improve relationships between people who do the same things.

There are certain individuals present in the scene who abstain from participating in the main group activity. Figure 6 depicts this scenario, with person IDs and actions

on the left and partial interaction weights on the right. Individuals traveling and queuing have lower interaction weights, suggesting less interaction. In order to reduce the possibility of misunderstanding between queuing and moving activity, this purposeful suppression of interactions stops messages from spreading between moving and queuing throughout the graph inference process.

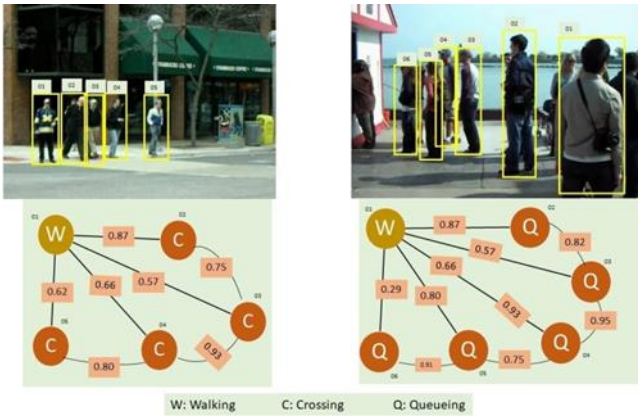

Fig. 6. Visualization of interactions. The degree of interaction between distinct actions is diminished.

9. Conclusion

Group activity recognition is the process of identifying the activity that a group is carrying out by looking at the actions and interactions of various group members. Conventional methodologies in this domain frequently employ handcrafted features to recognize group activities. Nevertheless, these techniques have difficulty capturing interdependencies and temporal dynamics among individuals. Applications in the real world have provided evidence of deep learning's efficacy in classifying group activities. Multiple relational graphs have been developed to represent various sorts of relationships and interactions among group members through the deep model. The GAR performance and computational speed were improved by introducing the deep learning model CNN and GCN. By incorporating a deformable convolution module into the CNN layer for individual action recognition, the deep model's performance is improved, achieving an accuracy of 93 : 52%. Finally, the GCN model is used to analyse a multiple-person relational graph model for the recognition of group activity, which improves the computational speed of GAR and enhances the accuracy to 95 : 4%. Two public datasets were used for experiments and comparisons. The results showed that modelling both the individual's appearance and the interaction context information can improve the accuracy of group activity recognition. Graph Convolutional Networks (GCNs) perform well in recognizing group activities by modelling complicated relationships among individuals. GCNs use graph structures to determine interactions and dependencies among group members, improving representational learning accuracy.

References

[1] G. J. Qi, H. Larochelle, B. Huet, J. Luo, and K. Yu., "Guest editorial: Deep learning for multimedia computing," in IEEE Transactions on Multimedia,, vol. 17, no. 11, pp. 1873–1874, 2015.

[2] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey. image and vision computing," in Image and Vision Computing, vol. 60, 2017.

[3] M. R. Amer, P. Lei, and S. Todorovic, "Hirf: hierarchical random field for collective activity recognition in videos," in Proceedings of the European Conference on Computer Vision, pp. 572-585, Zurich, Switzerland, 2014. Springer.

[4] Wongun Choi, K. Shahid, and S. Savarese, "What are they doing? : Collective activity classification using spatio-temporal relationship among people," in IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pp. 1282–1289, Kyoto, Japan, 2009.

[5] M. S. Qi, J. Qin, A. N. Li, Y. H. Wang, J. B. Luo, and L. Van Gool, "Stagnet: An attentive semantic rnn for group activity recognition," in Proceedings of the 15th European Conference on Computer Vision,, pp. 104–120, Munich, Germany, 2018. Springer,.

[6] Z. W. Deng, A. Vahdat, H. X. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in Proceedings of IEEE Conference on Computer Vision and Pattern Re- cognition, pp. 4772–4781, Las Vegas, USA, 2016.

[7] Ritu Basak Mohammed Shaheen Alam Jony Rashidul Hasan Nabil Tofayet Sultan, Nusrat Jahan, "Machine learning in cyberbullying detection from social-media image or screenshot with optical character recognition," in International Journal of Intelligent Systems and Applications (IJISA), vol. 15, no. 2, pp. 1-13, 2023.

[8] S. A. Vahora and N. C. Chauhan, "A comprehensive study of group activity recognition methods in video," in Indian Journal of Science and Technology, vol. 10, no. 23, pp.1–11, 2017.

[9] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," in Computer Vision and Image Understanding, vol. 117, no. 6, pp. 663–659, 2013.

[10] Zhang, Shugang, Zhiqiang Wei, Jie Nie, Lei Huang, Shuang Wang, and Zhen Li, "A review on human activity recognition using vision-based method" in

Journal of healthcare engineering, 2017.

[11] Rajesh P. Chinchewadi Tarun Jaiswal Sushma Jaiswal, Harikumar Pallthadka, "Optimized image captioning: Hybrid transformers vision transformers and convolutional neural networks: Enhanced with beam search" in International Journal of Intelligent Systems and Applications (IJISA), vol. 16, no. 2, pp. 53–61, 2024.

[12] C. Fauzi and S. Sulistyo, "A survey of group activity recognition in smart building," in IEEE Proceedings of International Conference on Signals and Systems, pp. 13–19, Bali,, 2018.

[13] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review. acm computing surveys," in Journal of healthcare engineering, vol. 43, no. 3, 2011.

[14] S. A. Vahora and N. C. Chauhan, "A comprehensive study of group activity recognition methods in video," in Indian Journal of Science and Technology, vol. 10, no. 23, 2017.

[15] Wu Li-Fang, Qi Wang, Meng Jian, Yu Qiao, and Bo-Xuan Zhao, "A comprehensive review of group activity recognition in videos," in International Journal of Automation and Computing, vol. 18, no. 3, pp. 334–350, 2021.

[16] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 8, pp. 1549–1562, 2012.

[17] Ryoo MS and Aggarwal J., "Stochastic representation and recognition of high-level group activities," in International journal of Computer Vision, vol. 93, 2011.

[18] P. Prabhu S. Anthonisamy, "Human activity detection using profound learning with improved convolutional neural networks," in International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 21 pp. 606–616, 2024.

[19] S. S. Kulkarni and S Jadhav, "Insight on human activity recognition using the deep learning approach," in 2023 International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 1–5. IEEE, 2023.

[20] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 4325–4334, Honolulu, USA,, 2017.

[21] M. S. Ibrahim, S. Muralidharan, Z.W. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1971–1980, Las Vegas, USA, 2015.

[22] L. H. Lu, H. J. Di, Y. Lu, L. Zhang, and S. Z. Wang, "A two-level attention-based interaction model for multiperson activity recognition," in Neurocomputing,, vol. 322, 2018.

[23] M. S. Ibrahim and G. Mori, "Hierarchical relational networks for group activity recognition and retrieval," in Proceedings of the 15th European Conference on Computer Vision, pp. 742–758, 2018.

[24] Wang L., Xiong Y., Y. Wang Z., Qiao, and X.and Van Gool Lin, D. and Tang, "Temporal segment networks: Towards good practices for deep action recognition," in European conference on computer vision, pp. 20–36, Cham, 2016. Springer.

[25] X. Li and M. C. Chuah. Sbgar, "Semantics based group activity recognition," in Proceedings of IEEE International Conference on Computer Vision, pp. 2895–2904, Venice, Italy, 2017.

[26] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks.

[27] M. S. Wang, B. B. Ni, and X. K. Yang, "Recurrent modeling of interaction context for collective activity recognition" in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 7408–7416, Honolulu, USA,, 2017. IEEE,.

[28] J. C. Wu, L. M. Wang, L. Wang, J. Guo, and G. S. Wu, "Learning actor relation graphs for group activity recognition" in Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9956–9966, Long Beach, USA, 2019.

[29] S. M. Azar, M. G. Atigh, A. Nickabadi, and A. Alahi, "Convolutional relational machine for group activity recognition," in Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7884–7893, Long Beach, USA, 2018.

[30] Yao L., Liu Y., and Huang, "Spatio-temporal information for human action recognition," in EURASIP Journal on Image and Video Processing, vol. 39, pp. 1–9, 2016.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and L. Polosukhin, "Attention is all you need," in Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 6000–6010, 2017.

[32] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Tim Lillicrap, "A simple neural network module for relational reasoning," in NIPS, 2017.

[33] Zijian Kuang and Xinran Tie, "Improved actor relation graph based group activity recognition," 2020.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in NeurIPS, 2017.

[35] Choi W., Shahid K., and Savarese S., "Learning context for collective activity recognition," in Proceedings of the 24th IEEE Conference on Computer Vision and Pattern Recognition, pp. 3273–3280, Colorado, 2011.

[36] P. Z. Zhang, Y. Y. Tang, J. F. Hu, and W. S. Zheng, "Fast collective activity recognition under weak supervision," in IEEE Transactions on Image Processing, vol. 29, 2019.

[37] D. Wang W. Ouyang X. Zhu, Y. Zhou and R. Su., "Mlst-former: Multilevel spatial-temporal transformer for group activity recognition," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 7, pp.3383–3397, 2023.

[38] C. Yuan Q. Tian R. Yan, X. Shu and J. Tang, "Position-aware participation-contributed temporal dynamic model for group activity recognition," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp.7574–7588, 2022.