Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.074115

ARTICLE

Group Activity Recognition in Crowded Scenes Using Multi-Stage Feature Optimization and ST-GCN-LSTM Networks

Mohammed Alnusayri[1], Tingting Xue[2,3], Saleha Kamal[4], Nouf Abdullah Almujally[5], Khaled Alnowaiser[6], Ahmad Jalal[4,7,*] and Hui Liu[3,8,9,*]

[1]Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakaka, Saudi Arabia
[2]School of Environmental Science & Engineering, Nanjing University of Information Science and Technology, Nanjing, China
[3]Cognitive Systems Lab, University of Bremen, Bremen, Germany
[4]Department of Computer Science, Air University, Islamabad, Pakistan
[5]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[6]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia
[7]Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul, Republic of Korea
[8]Jiangsu Key Laboratory of Intelligent Medical Image Computing, School of Future Technology, Nanjing University of Information Science and Technology, Nanjing, China
[9]Guodian Nanjing Automation Co., Ltd., Nanjing, China

*Corresponding Authors: Ahmad Jalal. Email: ahmadjalal@mail.au.edu.pk; Hui Liu. Email: hui.liu@uni-bremen.de

Received: 02 October 2025; Accepted: 23 December 2025; Published: 08 May 2026

ABSTRACT: Group activity recognition in public environments is challenging due to dynamic formations, complex inter-person interactions, and frequent occlusions. Existing methods often emphasize individual actions, overlooking collective behavioral patterns. This work introduces a multi-modal framework integrating silhouette-based appearance and skeleton-based pose information for robust recognition in surveillance scenarios. You Only Look Once v11 (YOLOv11) detects persons, Segmenting Objects by LOcations version 2 (SOLOv2) segments instances, and AlphaPose extracts skeletons, followed by hierarchical grouping to form spatially coherent clusters. A hybrid feature extraction strategy combines handcrafted descriptors (Extended GIST (ExGIST), Distance Transform, Binary Robust Independent Elementary Features (BRIEF), Ridge) with deep representations, fused via multi-head attention. Feature selection is refined through a three-stage pipeline of Kernel Principal Component Analysis (K-PCA), mutual information ranking, and genetic algorithm-based optimization. Spatio-Temporal Graph Convolution Networks (ST-GCN) models spatio-temporal dependencies, while Long Short-Term Memory (LSTM) captures long-term dynamics for activity classification. On the Collective Activity Dataset (CAD), the framework achieves 96.80% accuracy, surpassing state-of-the-art approaches. Its modular design ensures scalability and adaptability for intelligent surveillance and smart city applications.

KEYWORDS: BRIEF features; YOLO v11; STGCN; group activity recognition; LSTM

1 Introduction

Many applications such as surveillance and crowed managements require analyzing group activities, which is known as Group Activity Recognition (GAR) in computer vision. While individual activity recognition focusses on local actions of individuals, the analysis of group activities requires modeling

dependencies and interactions within the group. Such tracking faces many challenges such as occlusion, dynamic formations, and complex interactions.

Ongoing research has employed different deep learning approaches for GAR, which show promises. For example, graph convolutional networks (GCNs) are used for modeling human pose and interaction graphs [1] while the Spatio-Temporal Graph Convolutional Networks (ST-GCNs) is used to handle skeletal joint information and spatial relations [2]. Additionally, LSTM has approved its effectiveness in handling temporal dynamics [3]. Moreover, one of the advanced strategies that are used for feature extraction is multi-stage feature optimization [4] which allows to distinguish features at different spatial and temporal scales.

This research proposes a unified pipeline for group activity recognition in public settings, enhancing image quality via z-score normalization and bilateral filtering, selecting motion-relevant keyframes with a Convolutional Neural Network (CNN), detecting humans using YOLOv11, segmenting with SOLOv2, clustering via grouping algorithms, tracking identities with ByteTrack, and extracting skeletons via AlphaPose. Multi-modal features (ExGIST, Distance Transform, BRIEF, Ridge) are fused through multi-head attention, refined by a three-stage pipeline (Kernel PCA, mutual information ranking, genetic algorithm optimization), then modeled for spatio-temporal dependencies with ST-GCN and long-term dynamics with LSTM for classification. Validated on the Collective Activity Dataset, it exhibits robustness, scalability, and interpretability in real-world scenarios.

The novelty stems from its synergistic integration of appearance- and pose-based modalities via optimized fusion and a three-stage feature refinement of handcrafted/deep descriptors, enabling multi-head attention for cross-modal learning and resilience to occlusions/dynamic groups unlike single-modality or sequential approaches. The paper is structured as: Section 2 (GAR literature), Section 3 (framework/methodology), Section 4 (experiments/results), and Section 5 (conclusions/future work).

2 Literature Review

Group Activity Recognition (GAR) has gained significant research attention due to its relevance in surveillance, crowd management, and social behavior analysis. Over time, methodologies have evolved from focusing on individual actions to capturing complex group-level dynamics. Recent approaches can broadly be categorized into transformer and attention-driven frameworks, and multi-modal systems.

2.1 Transformer and Attention-Based Approach for GAR

Transformers and attention-based methods have become the dominant paradigm in group activity recognition, offering effective modeling of contextual cues and inter-actor dependencies. Du and Wang [5] introduced a contextual transformer network that integrates global scene tokens with localized features, while Li et al. [6] proposed Action Sequence Transformer (ASTFormer) with action-centric aggregation to address unbalanced spatio-temporal interactions. Tamura [7] designed efficient attention mechanisms for group-level feature extraction, and Tamura et al. [8] extended this with deformable attention to capture subgroup dynamics. More recently, Chen and Liu [9] presented a hierarchical attention query transformer that disentangles feature entanglement through distributed attention.

2.2 Graph Based Frameworks for GAR

Graph-based approaches emphasize relational reasoning between actors and frames. Li et al. [10] developed a graph-based residual aggregation network that models inter-individual differences through residual relations, whereas Wang et al. [11] introduced a temporal semantic sub-graph network that explicitly encodes inter-frame dependencies. Qin et al. [12] proposed CBAM-STGCN (Convolutional Block Attention

Module-based Spatial Temporal Graph Convolutional Network), which enhances the baseline ST-GCN by integrating spatial and channel attention modules through the convolutional block attention mechanism. This design enables adaptive weighting of key joints and feature channels, improving discriminative feature learning in skeleton-based recognition. These studies collectively highlight the potential of graph structures and attention mechanisms for spatio-temporal modeling in activity understanding.

Beyond transformers and graphs, alternative directions have also been explored. Chappa et al. [13] proposed a self-supervised framework that leverages spatio-temporal cooperative learning without requiring annotated labels, while Dutta et al. [14] presented a large-scale survey highlighting challenges such as occlusion and computational complexity and outlining future directions including zero-shot learning and multimodal integration.

3 Material and Methods

The proposed framework offers a unified pipeline for group activity recognition in public settings. It preprocesses images via z-score normalization and bilateral filtering for quality enhancement and uses a CNN-based keyframe selector to retain motion-relevant frames. Humans are detected with YOLOv11 and segmented via SOLOv2, clustered spatially with grouping algorithms, tracked identity-wise using ByteTrack, and posed with AlphaPose for joint dynamics. Multi-modal features (ExGIST, Distance Transforms, BRIEF, Ridge) capture motion, structure, and appearance, fused through multi-head attention and hierarchical integration. Features are refined via Kernel PCA, Mutual Information, Genetic Algorithm selection, and Recursive Feature Elimination. ST-GCN models spatio-temporal dependencies, while LSTM handles long-term dynamics for classification, ensuring robustness, scalability, and interpretability. Fig. 1 depicts the framework.

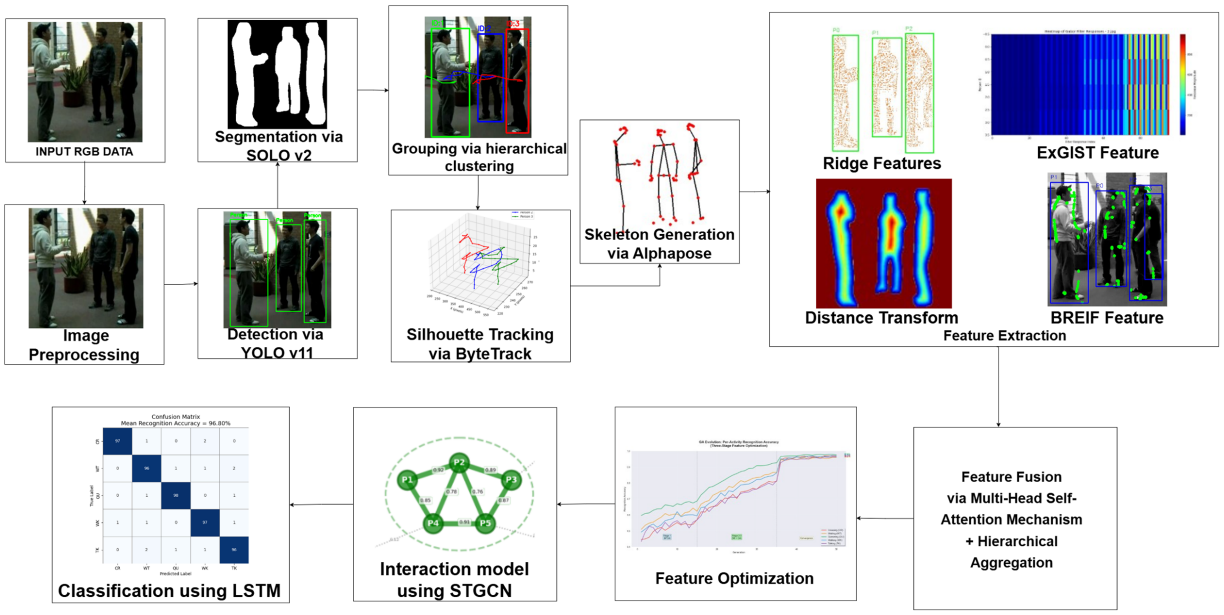


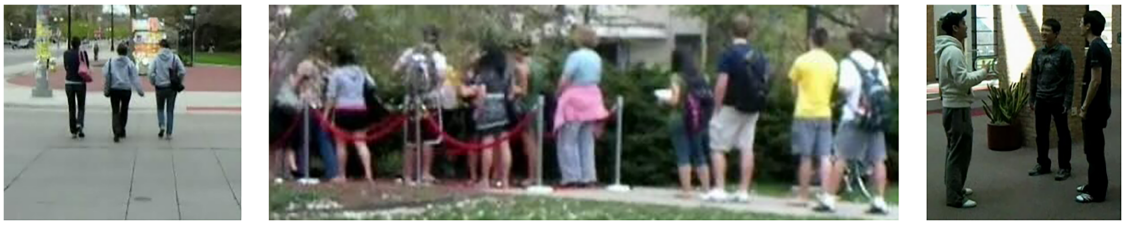
Figure 1: Architectural framework of the proposed model for GAR.

3.1 Preprocessing

To prepare the input video data for efficient and informative processing, a multi-step preprocessing pipeline is applied as showcased in Fig. 2, that includes normalization, noise removal, and motion-aware keyframe selection. Initially, each image frame is subjected to z-score normalization, a standardization technique used to normalize pixel intensities across frames. For a given frame I, the normalized image I' is computed as:

$$I'(x, y) = \frac{I(x, y) - \mu}{\sigma} \tag{1}$$

where μ and σ represent the mean and standard deviation of pixel intensities in the frame. This ensures consistent intensity distribution across the dataset and enhances the performance of downstream deep learning models. A non-local means bilateral filter is applied to suppress noise while preserving edges and textures essential for silhouette and motion detection.



(a) **(b)** **(c)**

Figure 2: Results of image preprocessing on (**a**) Crossing (**b**) Queueing (**c**) Talking on CAD.

Following normalization and denoising, motion-aware keyframe selection computes scores $M_t = \sum_{x,y} | F_t(x, y) - F_{t-1}(x, y) |$ for consecutive grayscale frames F_{t-1} and F_t.

Proportional allocation sets k_i keyframes per segment i (of S segments):

$$k_i = \left\lfloor T \cdot \frac{\sum_{t \in S_i} M_t}{\sum_{j=1}^{S} \sum_{t \in S_j} M_t} \right\rfloor \tag{2}$$

Clustering selects representatives: for each segment i, cluster frames into k_i groups $C_{i,j}$ with centers μ_j, then pick the closest frame per cluster. The keyframe set $\hat{K}$ is:

$$\hat{K} = \bigcup_{i=1}^{S} \{ \arg\min_{f \in C_{i,j}} \| f - \mu_j \|^2 \; j = 1, \ldots, k_i \} \tag{3}$$

This unified approach incorporates motion scoring, allocation, and selection to retain informative; representative keyframes while discarding redundancy.

3.2 Silhouette Detection

To localize human silhouettes in keyframes, the framework employs YOLOv11, a high-performance object detector optimized for speed and precision. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, the network partitions it into $S \times S$ cells, where each predicts B bounding boxes with parameters:

$$b_i = \left(x_i, y_i, w_i, h_i, C_i, {p_i}^{(1)}, {p_i}^{(2)}, \ldots, {p_i}^{(N)} \right) \tag{4}$$

here, (x_i, y_i) denote normalized box centers, (w_i, h_i) the dimensions, C_i the objectness score, and ${p_i}^{(j)}$ the class probabilities. For silhouettes, the detection score is:

$$Score_i = C_i \cdot {p_i}^{(Person)} \tag{5}$$

Detections below threshold τ are discarded, and Non-Maximum Suppression (NMS) retains boxes with the highest scores by eliminating overlaps exceeding IoU threshold θ:

$$IoU\left(b_m, b_j\right) = \frac{Area\left(b_m \cap b_j\right)}{Area\left(b_m \cup b_j\right)} > \theta \tag{6}$$

The final output is a set of high-confidence human silhouettes per frame. Fig. 3 illustrates few examples of silhouette detection.

(a) (b) (c)

Figure 3: Results of silhouette detection on (**a**) Walking (**b**) Queuing (**c**) Crossing on CAD.

3.3 Silhouette Segmentation

After detecting silhouettes, SOLOv2 is employed for pixel-level instance segmentation. Given an image $I \in \mathbb{R}^{H\times W\times 3}$, the model divides it into $S \times S$ grids. Each grid cell (i, j) predicts an object center and generates a dynamic convolution kernel K_{ij}, applied to the shared feature map F:

$$M_{ij} = \sigma\left(K_{ij} * F\right) \tag{7}$$

where $*$ denotes dynamic convolution and $\sigma(\cdot)$ normalizes pixel probabilities. Final mask is binarized as:

$$\hat{M}_{ij}(x, y) = \begin{cases} 1, & if\ M_{ij}(x, y) \geq \tau \\ 0, & otherwise \end{cases} \tag{8}$$

This anchor-free approach jointly optimizes category classification and mask prediction using a combined segmentation loss:

$$L_{seg} = \lambda_1 L_{cls} + \lambda_2 L_{dice} \tag{9}$$

yielding precise silhouette masks even under occlusion and overlap. Fig. 4 illustrates results.

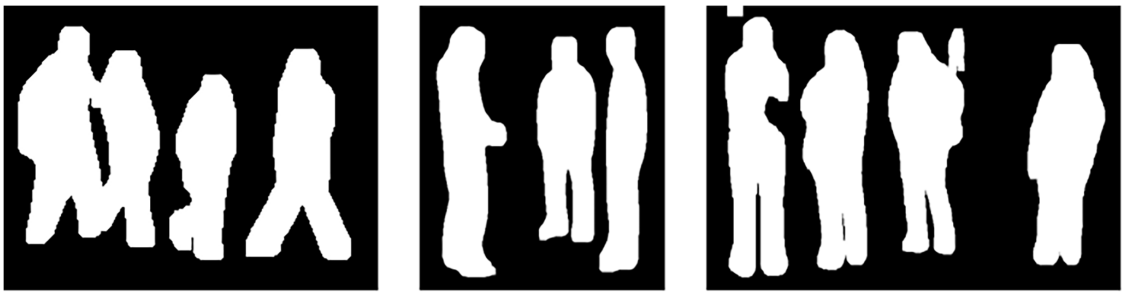


Figure 4: Results of silhouette segmentation on (**a**) Crossing (**b**) Talking (**c**) Crossing on CAD.

3.4 Silhouette Grouping

To identify spatially coherent interaction groups, a hierarchical clustering mechanism is applied to centroids derived from SOLOv2 instance masks as displayed in Fig. 5. For a binary mask $M(x, y)$, the centroid is computed from spatial moments:

$$(c_x, c_y) = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right) \tag{10}$$

$$m_{pq} = \sum_x \sum_y x^p y^q M(x, y) \tag{11}$$

Figure 5: Results of silhouette grouping on (**a**) Talking (**b**) Waiting (**c**) Walking on CAD.

Pairwise distances between individuals form a matrix $D \in \mathbb{R}^{n \times n}$:

$$D_{ij} = \|c_i - c_i\|_2 \tag{12}$$

Hierarchical clustering with Ward's linkage merges clusters A and B using:

$$\triangle(A, B) = \frac{|A||B|}{|A| + |B|} \|\overline{c}_A - \overline{c}_B\|^2 \tag{13}$$

The dendrogram is pruned with an adaptive threshold $T = \alpha \cdot h_{avg}$, ensuring scale robustness. Groups are retained only if they (i) contain >2 members, (ii) exhibit intra-group distances below T, and (iii) persist across frames, ensuring stable and behaviorally meaningful interaction clusters.

3.5 Silhouette Tracking

To preserve temporal coherence and individual identities, ByteTrack is adopted for multi-object tracking. Tracklets are first aligned with high-confidence detections using bounding box IoU:

$$IoU\left(B_t^i, D_t^j\right) = \frac{\left|B_t^i \cap D_t^j\right|}{\left|B_t^i \cup D_t^j\right|} \tag{14}$$

Occlusion handling is enhanced with a secondary stage that incorporates low-confidence detections, while MaskIoU from SOLOv2 masks improves precision:

$$MaskIoU\left(M_t^i, M_t^j\right) = \frac{\left|M_t^i \cap M_t^j\right|}{\left|M_t^i \cup M_t^j\right|} \tag{15}$$

Group stability is validated over a sliding window of N frames using a consistency score:

$$C\left(G\right) = w_1 S_s + w_2 M_s + w_3 I_d \tag{16}$$

where S_s, M_s, and I_d capture spatial stability, membership consistency, and interaction duration. Groups below threshold are discarded, while state transitions manage formation, merging, splitting, and dissolution. This ensures stable, interpretable group tracking in dynamic scenes. Fig. 6 shows silhouette tracking.

3.6 Skeleton and Keypoint Generation

To obtain a structured representation of human pose from Red, Green and Blue (RGB)-based silhouette regions, Alpha Pose was utilized for precise keypoint estimation. Alpha Pose is a top-down, multi-person pose estimator known for its high accuracy in crowded scenes. It first detects individuals using a bounding-box-based detector and then applies a single-person pose estimator to each detected region, making it suitable for the previously segmented and grouped silhouettes in the current pipeline.

Given an input RGB image $I \in \mathbb{R}^{H \times W \times 3}$, Alpha Pose first applies a human detector $D\left(\cdot\right)$, yielding a set of bounding boxes $\mathcal{B} = \{B_1, B_2, \ldots, B_N\}$, where each $B_i \in \mathbb{R}^4$ defines the spatial extent of an individual.

For each bounding box B_i, a pose estimation model $P\left(\cdot\right)$ predicts a set of 2D key points $\mathcal{K}_i = \{k_1, k_2, \ldots, k_J\}$, where $k_j = \left(x_j, y_j, c_j\right)$, representing the pixel coordinates and confidence score of joint j, and J is the total number of skeletal joints (typically 17 in the COCO format). The overall skeleton map S for frame t is thus defined as:

$$S_t = \bigcup_{i=1}^{N} \mathcal{K}_i^t \; where \; \mathcal{K}_i^t = \left(B_i^t\right) \tag{17}$$

To ensure that only high confidence key points are retained, a thresholding mechanism is applied:

$$k_j \in \mathcal{K}_i \; iff \; c_j \geq \tau \tag{18}$$

where τ is the confidence threshold, empirically set to 0.6 for optimal balance between noise suppression and joint retention. Each skeleton is then aligned with the tracked silhouette identity using spatial correspondence, ensuring consistent mapping of pose features over time. This step forms the foundation for the motion-aware feature extraction that follows. Fig. 7 illustrates skeleton generation.

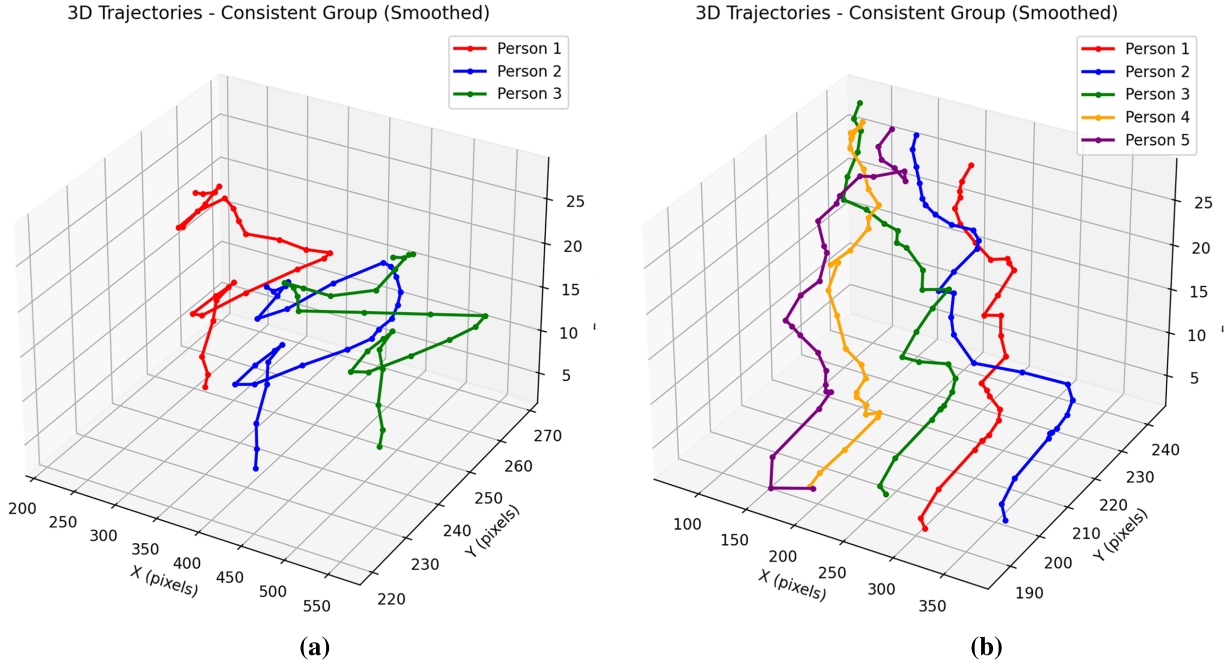


Figure 6: Results of silhouette tracking on (**a**) Talking (**b**) Queuing on CAD.

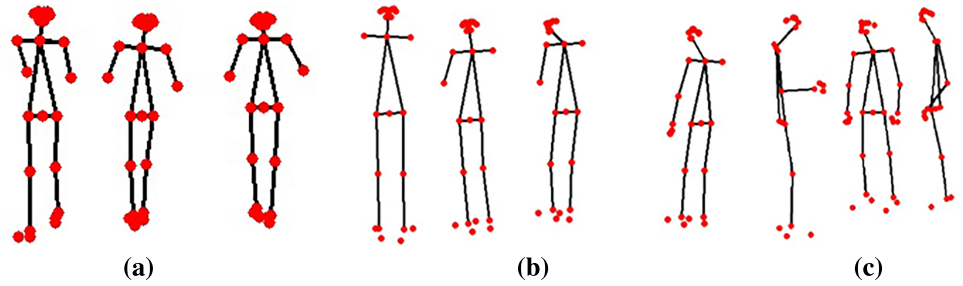


Figure 7: Results of skeleton generation on (**a**) Crossing (**b**) Waiting (**c**) Talking on CAD.

3.7 Feature Extraction

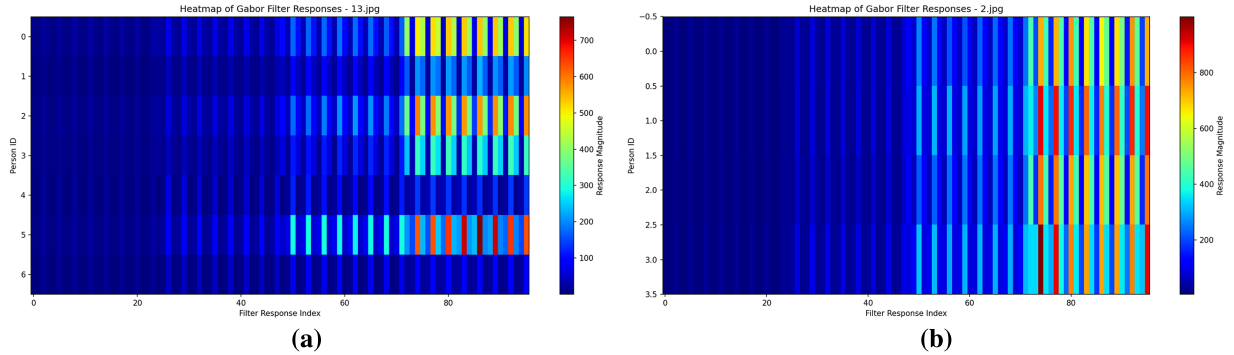
3.7.1 ExGIST Features

ExGIST features are extended gradient-based descriptors that encode the spatial distribution of edges and orientations within a silhouette. Unlike simple gradient histograms, ExGIST captures multi-scale directional information, making it effective for representing both local body structures and global group formations. The descriptor is computed by partitioning the silhouette image into a fixed grid and calculating

oriented gradient energy in each cell, followed by concatenation into a high-dimensional vector:

$$ExG = \bigcup_{i=1}^{N} \left(\sum_{x,y \in C_i} \nabla I(x,y) \cdot \theta \right) \tag{19}$$

where C_i denotes the i-th cell, $\nabla I(x,y)$ represents the gradient magnitude, and θ is the gradient orientation. This structured representation encodes edge alignment and regional intensity variations, enabling discrimination between subtle group-level postures and motion patterns. By preserving both orientation consistency and spatial hierarchy, ExGIST features enhance recognition of complex activities such as standing in formation or conversational clustering. Fig. 8 shows extracted ExGIST features.



(a) **(b)**

Figure 8: Results of ExGIST on (**a**) Crossing (**b**) Waiting on CAD.

3.7.2 Distance Transform

To capture the spatial structure and boundary proximity of silhouettes, the distance transform was utilized. This transformation is particularly effective in preserving shape-related cues by computing the minimal Euclidean distance of each foreground pixel to the nearest background pixel within a binary silhouette mask. Given a binary silhouette image $S(x,y)$, where $S(x,y) = 1$ represents foreground (silhouette) and $S(x,y) = 0$ represents background, the Euclidean distance transform $D(x,y)$ is defined as:

$$D(x,y) = \min_{(x',y') \in \mathcal{B}} \sqrt{(x-x')^2 + (y-y')^2} \tag{20}$$

where $\mathcal{B}$ is the set of all background pixels. This results in a smooth map where each pixel encodes its proximity to the silhouette boundary.

The generated distance map highlights contour-preserving details and internal structure, which are particularly useful for understanding pose deformation, limb articulation, and group cohesion. Furthermore, applying normalization to the distance map ensures scale invariance, making the features robust across varying body sizes and camera distances. These distance-based features are later vectorized and concatenated with other descriptors, enriching the feature space with spatial configuration cues that are critical for interaction analysis. Fig. 9 shows distance transform features.

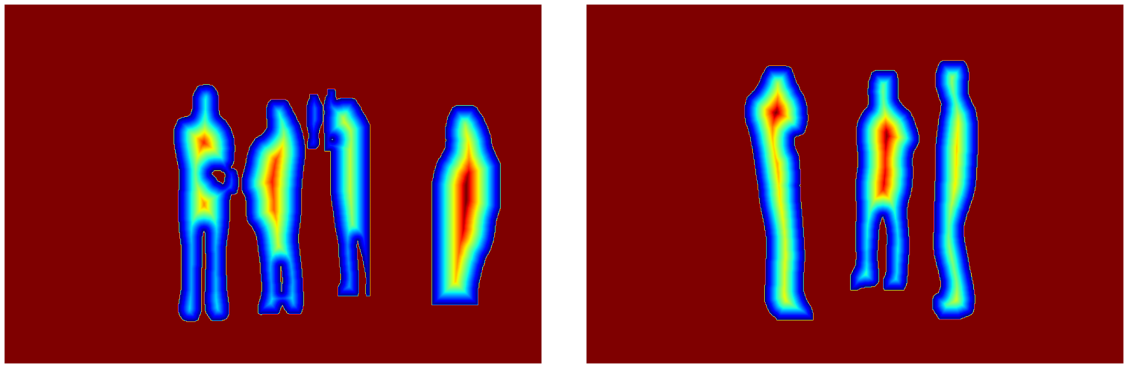
3.7.3 BRIEF

To encode compact local descriptors from segmented silhouette regions, the BRIEF descriptor was employed due to its computational efficiency and high discriminative capacity. BRIEF operates by comparing pixel intensities within a predefined smoothed image patch and encoding the results as binary strings,

thus capturing key texture patterns and fine-grained spatial variations. For each keypoint p located in the silhouette region, a binary descriptor $f_p \in \{0,1\}^n$ is computed using a set of n pairwise pixel comparisons:

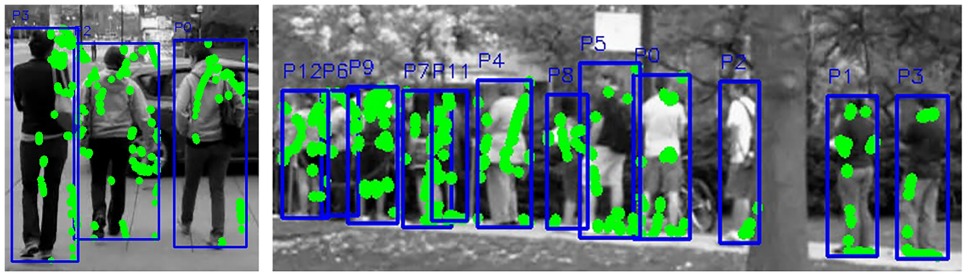
$$f_p[i] = \begin{cases} 1, & if\ I(p+x_i) < I(p+y_i) \\ 0, & otherwise \end{cases} \quad for\ i = 1, 2, \ldots, n \tag{21}$$

where x_i and y_i are predefined sampling offsets within the local patch around p, and I denotes the pixel intensity. This binary encoding yields a low-dimensional yet distinctive feature representation that is highly robust to noise, scale variations, and minor deformations. In the context of silhouette-based human activity recognition, BRIEF enables efficient encoding of motion boundaries, limb orientations, and texture contrast without requiring gradient computation. The resulting descriptors from all key points within a frame are aggregated using a bag-of-words scheme, forming a global binary histogram that contributes to the overall feature fusion pipeline. Fig. 10 shows BRIEF features.



(a) **(b)**

Figure 9: Results of distance transform on (**a**) Waiting (**b**) Talking on CAD.



(a) **(b)**

Figure 10: Results of BRIEF on (**a**) Crossing (**b**) Queuing on CAD.

3.7.4 Ridge Features

Ridge features are structural descriptors that capture elongated intensity patterns within silhouettes or body contours, effectively encoding the orientation and continuity of motion trajectories. In group activity recognition, these features help identify coordinated behaviors where multiple individuals exhibit aligned or parallel movements. Mathematically, ridge extraction is based on the second-order derivatives of the image

intensity I_{xy}, where ridge points correspond to local maxima of the eigenvalues of the Hessian matrix:

$$H(x, y, \sigma) = \begin{vmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{vmatrix} \tag{22}$$

here, I_{xx}, I_{yy}, and I_{xy} denote the second-order partial derivatives. Strong eigenvalue responses indicate ridge-like structures, which are then aggregated into histograms to preserve orientation distributions. By emphasizing directional continuity, ridge features contribute discriminative cues that complement motion-based descriptors, particularly in scenarios involving group alignment such as walking or queue formation. Fig. 11 shows results of ridge feature.

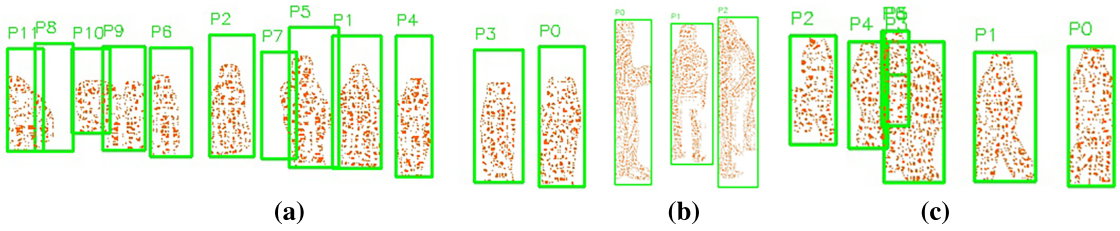


Figure 11: Results of ridge on (**a**) Crossing (**b**) Waiting (**c**) Talking on CAD.

3.8 Feature Fusion

A hybrid feature fusion strategy is employed to integrate descriptors derived from ExGISTs, distance transforms, BRIEF, and RIDGE. This approach leverages multi-head attention mechanisms and hierarchical aggregation to consolidate complementary spatial, temporal, and appearance-based information while maintaining contextual dependencies. Initially, each feature vector $F_i \in \mathbb{R}^d$ from modality i is passed through a multi-head self-attention module, which captures inter-feature relationships by computing query (Q), key (K), and value (V) projections:

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\intercal}}{\sqrt{d_k}}\right)v \tag{23}$$

where $Q = F_i W^Q$, $K = F_i W^K$, and $V = F_i W^V$, with W^Q, W^K, W^V being the trainable projection matrices and d_k the dimensionality of the key vectors. The multi-head mechanism allows the model to attend to information from different representation subspaces simultaneously, enhancing the expressiveness of the fused embedding.

Following attention-based refinement, the outputs from all modalities are hierarchically aggregated. This hierarchical fusion involves progressive concatenation and dimensionality reduction using fully connected layers to integrate high-level and low-level cues:

$$F_{fused} = \phi\left(W_h\left[F_1^{'} \parallel F_2^{'} \parallel \ldots \parallel F_n^{'}\right] + b_h\right) \tag{24}$$

where $\parallel$ denotes concatenation, $\phi(\cdot)$ is a non-linear activation, and W_h, b_h represent the fusion weights and bias. This approach ensures that redundant information is suppressed while salient interactions among features are preserved. The combined representation F_{fused} encapsulates motion dynamics, shape cues, and structural keypoint relationships, serving as a compact and discriminative input to subsequent optimization and classification stages.

3.9 Feature Optimization

To refine the fused representation F_{fused} and enhance its discriminative capacity while reducing computational complexity, a three-stage feature optimization strategy was adopted, integrating Kernel Principal Component Analysis (kPCA), Mutual Information (MI) ranking, and Genetic Algorithm (GA)-based selection. The objective of this stage is to suppress redundant and noisy components while retaining the most informative patterns for subsequent spatio-temporal modeling.

In the first stage, kPCA projects $F_{fused} \in \mathbb{R}^d$ into a non-linear manifold space defined by a kernel function $k(.,.)$, enabling the preservation of complex inter-feature relationships. Given the kernel matrix $K_{ij} = k\left(F_i, F_j\right)$, the centered kernel $\tilde{K}$ is decomposed as:

$$\tilde{K}\alpha = \lambda\alpha \tag{25}$$

where λ and α are the eigenvalues and eigenvectors, respectively. The top m components corresponding to the largest λ values are retained, producing the reduced representation $Z \in \mathbb{R}^m$. In the second stage, MI ranking quantifies the dependency between each feature Z_j and the class label Y using:

$$I\left(Z_j; Y\right) = \sum_{z_j} \sum_{y} p\left(z_j, y\right) \log \frac{p\left(z_j, y\right)}{p\left(z_j\right) p\left(y\right)} \tag{26}$$

Features with higher MI scores are prioritized, and a subset S_0 containing the top-ranked components is passed to the final optimization stage. In the third stage, a GA-based search identifies the optimal feature subset S^* that balances classification performance and compactness. Each chromosome encodes a binary selection mask over S_0, and the fitness function is defined as:

$$F\left(c\right) = \alpha \cdot Acc_{val}\left(c\right) - \beta \cdot \frac{\|c\|_1}{|S_0|} \tag{27}$$

where $Acc_{val}\left(c\right)$ is the validation accuracy using features indicated by mask c, and α, β control the trade-off between accuracy and subset size. This hierarchical optimization pipeline ensures that only the most relevant, non-redundant features are forwarded to the ST-GCN interaction model, leading to improved generalization, reduced overfitting, and lower inference costs. Fig. 12 shows feature optimization graph.

3.10 Interaction Model

To recognize group activities in public spaces, the optimized feature set F_{opt} is processed through a Spatial-Temporal Graph Convolutional Network (ST-GCN) that models dynamic inter-personal relationships. Each individual is represented by a skeletal graph with nodes as joints and edges capturing spatial bone connections and temporal motion continuity. The group interaction is formulated as a spatio-temporal graph $G = \left(V, E_s \bigcup E_t\right)$, where V denotes all joints, A_s encodes intra- and inter-person spatial links, and A_t captures temporal trajectories. The graph convolution at layer l is expressed as:

$$H^{(l+1)} = \sigma\left(\sum_{k=1}^{K} \tilde{A}_k H^{(l)} W_k\right) \tag{28}$$

where $\tilde{A}_k = A_k^{1/2} D_k^{-1/2}$ is the normalized adjacency matrix for partition k (of K spatial-temporal partitions), D_k is the degree matrix, W_k are learnable weights, and $\sigma(\cdot)$ is a non-linear activation (e.g., ReLU). The ST-GCN models groups as spatio-temporal graphs, capturing individual postures and collective dynamics. By integrating motion, geometry, and spatial relations, it distinguishes similar activities and generates robust embeddings for accurate group activity classification.

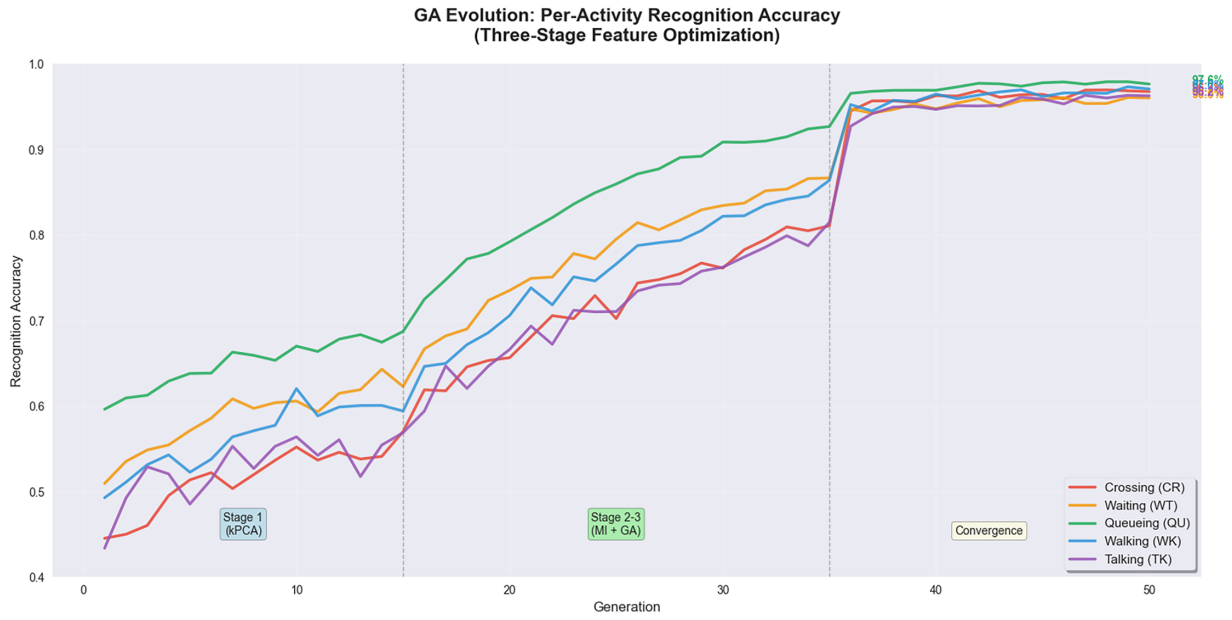


Figure 12: Graph of feature optimization on CAD.

3.11 Classification

The refined interaction-aware embeddings obtained from the ST-GCN are classified into predefined group activity categories using a LSTM network. LSTM networks are a variant of recurrent neural networks (RNNs) specifically designed to capture long-range temporal dependencies by mitigating the vanishing gradient problem through a gated memory mechanism. This makes them well-suited for recognizing group activities that evolve over extended time sequences. Given a sequence of feature vectors $\{x_t\}_{t=1}^{T}$ from the feature refinement stage, the LSTM updates its hidden state h_t and cell state c_t at each time step t according to:

$$i_t = \sigma\left(W_i x_t + U_i h_{t-1} + b_i\right) \tag{29}$$

$$f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right) \tag{30}$$

$$o_t = \sigma\left(W_o x_t + U_o h_{t-1} + b_o\right) \tag{31}$$

$$\tilde{c}_t = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \tag{32}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{33}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{34}$$

where i_t, f_t, and o_t denote the input, forget, and output gates, respectively; σ is the sigmoid activation; and $\odot$ represents element-wise multiplication. These gates regulate the flow of information, enabling the network to retain relevant temporal patterns while discarding irrelevant or redundant dynamics.

The final hidden state h_T encodes the temporal evolution of group interactions over the entire sequence. This representation is passed through a fully connected layer with a SoftMax activation function to produce class probabilities:

$$\hat{y} = softmax\left(W_{cls} h_T + b_{cls}\right) \tag{35}$$

where W_{cls} and b_{cls} are the classification weights and bias. By leveraging the temporal memory capability of LSTMs, the classifier captures nuanced variations in motion patterns and inter-person relationships that distinguish activities such as group walking, queuing, or conversing. This ensures that the final recognition decision reflects both the spatial structure and temporal progression of the group's behavior in real-world public space scenarios.

4 Experiments and Results

In this section, we present the experimental procedures conducted as part of this study and the corresponding results obtained to showcase the effectiveness and importance of the proposed model.

4.1 Datasets Description

To evaluate the effectiveness of the proposed system, experiments were conducted using the Collective Activity Dataset, a widely used benchmark for group activity recognition.

Collective Activity Dataset

CAD is a benchmark for group-level human activity recognition, containing 44 outdoor video clips (720 × 480 resolution) across five activity classes: crossing, waiting, queueing, walking, and talking.

4.2 Performance Measurement and Result Analysis

The proposed framework was evaluated on the Collective Activity Dataset and validated on the Volleyball Dataset to assess generalization in complex group activities. Key metrics included precision, recall, F1-score, confusion matrices, Receiver Operating Characteristic (ROC) curves, ablation studies, and SOTA comparisons. Tables 1 and 2 present class-wise confusion matrices, highlighting strong discriminative performance; Fig. 13 shows stable per-class precision, recall, and F1 trends; and Fig. 14's ROC curve confirms an optimal true/false positive balance, underscoring the model's robustness.

Table 1: CAD dataset classification accuracies (confusion matrix).

Classes	CR	WT	QU	WK	TK
CR	97	1	0	2	0
WT	0	96	1	1	2
QU	0	1	98	0	1
WK	1	1	0	97	1
TK	0	2	1	1	96
Mean Recognition Accuracy = 96.80%					

Note: CR = Crossing, WT = Waiting, QU = Queueing, WK = Walking, TK = Talking.

Table 2: Volleyball dataset classification accuracies (confusion matrix).

Classes	Attack	Defense	Transition	Set Piece	Goal	Card	Substitution	Other
Attack	96	2	1	0	1	0	0	0
Defense	3	95	1	0	0	0	0	1
Transition	1	2	95	1	0	0	0	1
Set Piece	0	0	1	96	2	0	0	1
Goal	1	0	0	2	96	1	0	0

(Continued)

Table 2 (continued)

Classes	Attack	Defense	Transition	Set Piece	Goal	Card	Substitution	Other
Card	0	0	0	0	1	97	1	1
Substitution	0	0	0	0	0	1	98	1
Other	0	1	1	0	0	0	0	98
Mean Recognition Accuracy = 96.38%								

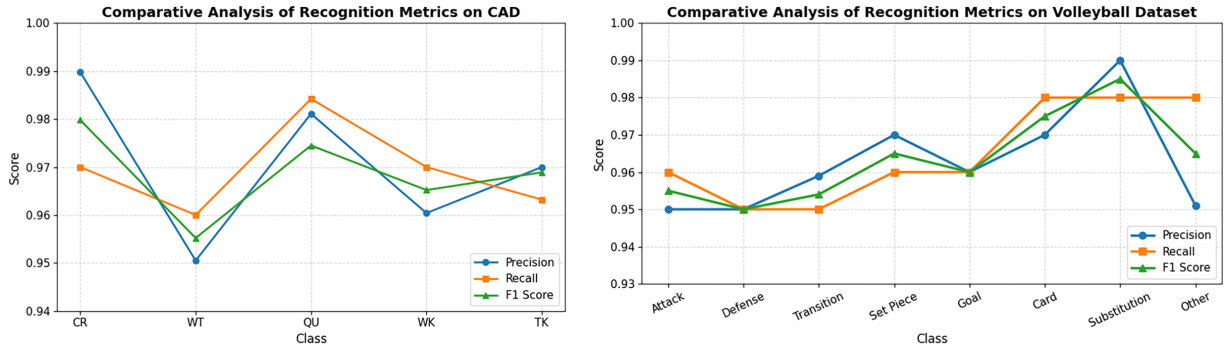


Figure 13: Comparative analysis of recognition metrics on (**a**) CAD (**b**) Volleyball dataset.

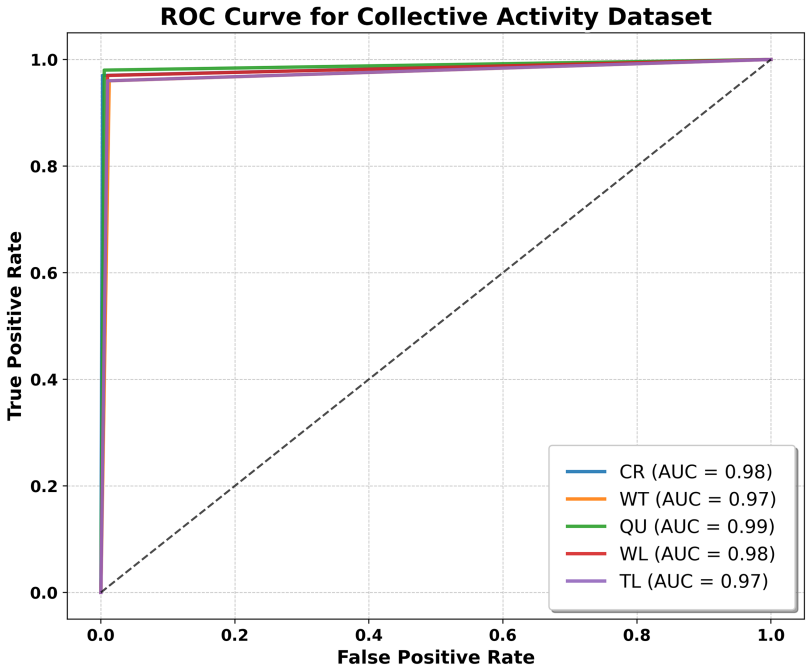


Figure 14: ROC curve for CAD.

Finally, a SOTA comparison is provided in Table 3, where the proposed method achieves superior performance compared to existing group activity recognition approaches.

Table 3: Comparison of classification accuracies with other state-of-the-art methods on CAD and volleyball dataset.

Author/Method	Collective Activity Dataset	Volleyball Dataset
Kim et al. [15]	81.50%	–
Ehsanpour et al. [16]	89.40%	–
Trehan and Aakur [17]	90.41%	–
Kong et al. [18]	90.60%	–
Gavrilyuk et al. [19]	90.80%	89.30%
Yan et al. [20]	92.50%	–
Li et al. [21]	93.60%	–
Han et al. [22]	–	93.20%
Wang & Mohamed [23]	–	94.10%
Zahra et al. [24]	–	94.50%
Proposed	**96.80%**	**96.38%**

4.3 Ablation Study

An ablation study on the Collective Activity Dataset assessed each module's contribution to the framework. The full pipeline achieved 96.80% accuracy, with removals causing declines: excluding silhouette segmentation and skeleton generation dropped it to 92% (key for group interactions); omitting ExGIST and Ridge Descriptors reduced it to 95% (vital for motion/appearance cues); excluding feature fusion or optimization yielded 93% and 94%, respectively (essential for integration/refinement). Table 4 confirms each module's significance, with the full combination optimal.

Table 4: Ablation study of proposed framework.

Exp	PR	SD	SS	ST	SG	SK	EG	DT	BR	RG	FF	FO	CAD
FM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	96.80
w PR	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	96
w SD	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	94
w SS	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	92
w ST	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	95
w SG	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	96
w SK	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	92
w EG	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	95
w DT	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	94
w BR	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	96
w RG	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	95
w FF	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	93
w FO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	94

Note: FM = Full Model, w = Without, PR = Preprocessing, SD = Silhouette Detection, SS = Silhouette Segmentation, ST = Silhouette Tracking, SG = Silhouette Grouping, SK = Skeleton Keypoints Generation, EG = ExGIST, DT = Distance Transform, BR = BRIEF Descriptor, RG = RIDGE Features, FF = Feature Fusion, FO = Feature Optimization, CAD = Collective Activity Dataset.

5 Conclusion

This work introduces a multi-modal framework that fuses silhouette-based appearance and skeleton-based pose features for robust group activity recognition in surveillance. It detects humans with YOLOv11 and segments them via SOLOv2, followed by ByteTrack for tracking and AlphaPose for pose estimation. Multi-modal features (ExGIST, Distance Transform, BRIEF, Ridge) are extracted, refined through a three-stage pipeline (Kernel PCA, mutual information ranking, genetic algorithm optimization), and fused via multi-head attention. Spatio-temporal dependencies are modeled with ST-GCN, and long-term dynamics with LSTM for classification. Evaluated on the Collective Activity Dataset, it achieves 96.80% accuracy, highlighting its potential for intelligent surveillance and smart cities.

Despite strong performance, limitations include untested generalization to cross-domain scenarios (e.g., varying lighting, viewpoints, or densities), higher computational costs in crowded scenes, and lack of support for real-time online streams. Future efforts will address domain adaptation, scalable dense-crowd processing, and edge/cloud-optimized variants.

Acknowledgement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Funding Statement: The APC was funded by the Open Access Initiative of the University of Bremen and the DFG via SuUB Bremen. This research is supported and funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: study conception and design: Saleha Kamal, Mohammed Alnusayri and Tingting Xue; data collection: Saleha Kamal, Khaled Alnowaiser and Nouf Abdullah Almujally; analysis and interpretation of results: Mohammed Alnusayri, Khaled Alnowaiser and Ahmad Jalal; draft manuscript preparation: Saleha Kamal, Hui Liu and Ahmad Jalal. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All publicly available datasets are used in the study.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tang Y, Lu J, Wang Z, Yang M, Zhou J. Learning semantics-preserving attention and contextual interaction for group activity recognition. IEEE Trans Image Process. 2019;28(10):4997–5012. doi:10.1109/TIP.2019.2914577.
2. Li Z, Chang X, Li Y, Su J. Skeleton-based group activity recognition via spatial-temporal panoramic graph. In: Proceedings of the Computer Vision—ECCV 2024; 2024 Sep 29–Oct 4; Milan, Italy. p. 252–69.
3. Wang C, Mohmamed ASA, Bin Mohd Noor MH, Yang X, Yi F, Li X. ARN-LSTM: a multi-stream fusion model for skeleton-based action recognition. arXiv:2411.01769. 2024.
4. Ibrahim MS, Muralidharan S, Deng Z, Vahdat A, Mori G. A hierarchical deep temporal model for group activity recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. Piscataway, NJ, USA: IEEE; 2016. p. 1971–80.
5. Du Z, Wang Q. Exploring global context and position-aware representation for group activity recognition. Image Vis Comput. 2024;149(6):105181. doi:10.1016/j.imavis.2024.105181.
6. Li W, Yang T, Wu X, Du XJ, Qiao JJ. Learning action-guided spatio-temporal transformer for group activity recognition. In: Proceedings of the 30th ACM International Conference on Multimedia; 2022 Oct 10–14; Lisboa, Portugal. p. 2051–60.
7. Tamura M. Design and analysis of efficient attention in transformers for social group activity recognition. Int J Comput Vis. 2024;132(10):4269–88. doi:10.1007/s11263-024-02082-y.

8. Tamura M, Vishwakarma R, Vennelakanti R. Hunting group clues with transformers for social group activity recognition. In: Proceedings of the Computer Vision—ECCV 2022; 2022 Oct 23–27; Tel Aviv, Israel. p. 19–35.
9. Chen X, Liu Z. Hierarchical query design and distributed attention in transformer for player group activity recognition in sports analysis. Sci Rep. 2025;15(1):31571. doi:10.1038/s41598-025-16752-5.
10. Li W, Yang T, Wu X, Yuan Z. Learning graph-based residual aggregation network for group activity recognition. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence; 2022 Jul 23–29; Vienna, Austria. p. 1102–8.
11. Wang D, Liu J, Zhou Y. Group activity recognition based on temporal semantic sub-graph network. In: Proceedings of the 2022 14th International Conference on Machine Learning and Computing (ICMLC); 2022 Feb 18–21; Guangzhou, China. p. 401–6.
12. Qin J, Zhang S, Wang Y, Yang F, Zhong X, Lu W. Improved skeleton-based activity recognition using convolutional block attention module. Comput Electr Eng. 2024;116:109231. doi:10.1016/j.compeleceng.2024.109231.
13. Chappa NVR, Nguyen P, Nelson AH, Seo HS, Li X, Daniel Dobbs P, et al. SoGAR: self-supervised spatiotemporal attention-based social group activity recognition. IEEE Access. 2025;13:33631–42. doi:10.1109/access.2025.3541986.
14. Dutta SJ, Boongoen T, Zwiggelaar R. Human activity recognition: a review of deep learning-based methods. IET Comput Vis. 2025;19(1):e70003. doi:10.1049/cvi2.70003.
15. Kim J, Lee M, Heo JP. Self-feedback DETR for temporal action detection. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France. p. 10252–62.
16. Ehsanpour M, Abedin A, Saleh F, Shi J, Reid I, Rezatofighi H. Joint learning of social groups, individuals action and sub-group activities in videos. In: Proceedings of the 16th European Conference on Computer Vision—ECCV 2020; 2020 Aug 23–28; Glasgow, UK. p. 177–95.
17. Trehan S, Aakur SN. Self-supervised multi-actor social activity understanding in streaming videos. In: Proceedings of the 27th International Conference on Pattern Recognition; 2024 Dec 1–5; Kolkata, India. p. 293–309.
18. Kong L, Pei D, He R, Huang D, Wang Y. Spatio-temporal player relation modeling for tactic recognition in sports videos. IEEE Trans Circuits Syst Video Technol. 2022;32(9):6086–99. doi:10.1109/TCSVT.2022.3156634.
19. Gavrilyuk K, Sanford R, Javan M, Snoek CGM. Actor-transformers for group activity recognition. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 836–45.
20. Yan R, Xie L, Tang J, Shu X, Tian Q. HiGCIN: hierarchical graph-based cross inference network for group activity recognition. IEEE Trans Pattern Anal Mach Intell. 2023;45(6):6955–68. doi:10.1109/TPAMI.2020.3034233.
21. Li S, Cao Q, Liu L, Yang K, Liu S, Hou J, et al. GroupFormer: group activity recognition with clustered spatial-temporal transformer. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 13648–57.
22. Han M, Zhang DJ, Wang Y, Yan R, Yao L, Chang X, et al. Dual-AI: dual-path actor interaction learning for group activity recognition. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 2980–9.
23. Wang C, Mohamed ASA. Group activity recognition in computer vision: a comprehensive review, challenges, and future perspectives. arXiv:2307.13541. 2023.
24. Zahra I, Wu Y, Alhasson HF, Alharbi SS, Aljuaid H, Jalal A, et al. Dynamic graph neural networks for UAV-based group activity recognition in structured team sports. Front Neurorobot. 2025;19:1631998. doi:10.3389/fnbot.2025.1631998.