Sensors and Materials, Vol. 38, No. 5 (2026) 2435–2453
MYU Tokyo

S & M 4446

Scalable Video Sensors for Understanding Multiperson Sports Behavior Recognition in Dynamic Scenes

Saleha Kamal,[1,2†] Yanfeng Wu,[1†] Nouf Abdullah Almujally,[3†]
Ahmad Jalal,[2,4*†] and Hui Liu[1,5,6**†]

[1]Guodian Nanjing Automation Co., Ltd., Nanjing, China
[2]Faculty of Computing and AI, Air University, E-9, Islamabad 44000, Pakistan
[3]Department of Information Systems, College of Computer and Information Sciences,
Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[4]Department of Computer Science and Engineering, College of Informatics,
Korea University, Seoul 02841, South Korea
[5]Jiangsu Key Laboratory of Intelligent Medical Image Computing, School of Artificial Intelligence,
School of Future Technology, Nanjing University of Information Science and Technology, Nanjing, China
[6]Cognitive Systems Lab, University of Bremen, Bremen 28359, Germany

(Received November 7, 2025; accepted March 5, 2026)

Keywords: multiperson human behavior recognition, group behavior analysis, segmentation, silhouette tracking, multihead attention, human behavior interaction, sports video classification

The surge in multiperson sports videos has increased the demand for automated behavior recognition systems capable of understanding group-level activities. Recognizing collective behaviors such as coordination in basketball or formations in volleyball remains challenging owing to occlusions, rapid motion, and long video sequences, while many existing approaches focus primarily on individual actions rather than collective understanding. In this study, we advance noncontact vision-based sensing through a computer vision pipeline that aggregates per-person cues into sequence-level descriptors for sport classification using the MultiSports dataset, comprising basketball, football, aerobics, and volleyball. The proposed framework integrates preprocessing with denoising, normalization, and motion-guided keyframe extraction; human representation using silhouette detection, tracking, and skeleton estimation; and a hybrid feature strategy combining deep learned representations with handcrafted motion and shape descriptors, which are fused and modeled to capture interaction dynamics, followed by hierarchical classification. Experimental results demonstrate an overall classification accuracy of 91.25% under 10-fold cross-validation, validating the effectiveness of the proposed approach. The main contributions include a motion-aware keyframe selection strategy for long-duration videos, a hybrid feature representation for group behavior modeling, and an efficient recognition framework, while current limitations related to reliance on RGB data and fixed viewpoints motivate future work on adaptive temporal modeling and multimodal sensing.

*Corresponding author: e-mail: ahmadjalal@au.edu.pk
**Corresponding author: e-mail: hui.liu@uni-bremen.de
†These authors contributed equally to this work.
https://doi.org/10.18494/SAM5999

1. Introduction

The rapid proliferation of multiperson sports videos, driven by wearable and ambient sensing technologies, has revolutionized sports analytics, surveillance, and broadcasting into data-rich domains for automated understanding.[1] Human behavior recognition (HBR) from noncontact RGB cameras enables the real-time extraction of a behavioral cue, such as team coordination in basketball or positional alignment in volleyball, supporting performance optimization, injury prevention, and immersive fan experiences.[2] However, challenges such as player occlusions, rapid motions, and long untrimmed sequences (>180 s) hinder traditional human activity recognition (HAR) methods, which focus on isolated actions rather than group dynamics.

Existing HAR approaches in sports leverage deep learning, such as two-stream CNNs or graph neural networks, which perform well on single-person or trimmed clips but struggle with multiperson, long-form videos owing to scalability and viewpoint sensitivity.[3] Models such as slow fast or video masked autoencoder (MAE), capture temporal evolution but neglect interperson relations, yielding suboptimal results (e.g., <70% average precision on MultiSports validation). Handcrafted features such as optical flow or pose moments provide interpretability but lack integration with transformers for relational reasoning. Addressing these gaps demands sensing pipelines that aggregate per-person representations into sequence-level descriptors, highlighting sport-specific signatures such as pose diversity and motion synchronization, without exhaustive annotations.[4]

In the paper, we introduce a pipeline for video-level sport classification on MultiSports clips using motion-aware keyframe selection and robust human representation. Silhouettes are detected, segmented, tracked, and converted into skeletons to handle occlusions effectively. Hybrid features capturing motion, edges, deformations, and poses are fused via multihead attention, refined through contrastive and social transformer learning, and hierarchically classified using a RecNN for final SoftMax prediction.

Contributions include (1) a scalable framework shifting from individual to collective representations for better interpretability and efficiency; (2) tailored fusion of handcrafted and learned features, achieving overall classification accuracy, 91.25% sport accuracy, and 78.5% frame-mAP on MultiSports (5–7% above baselines); and (3) insights into visual sensing for behavior understanding, enabling real-time sports applications. The remainder of this paper is organized as follows. In Sect. 2, we present reviews related to this work. Sect. 3 comprises descriptions of the techniques used, and we present the experiments in Sect. 4. In Sect. 5, we present discussion limitations and future work. Lastly, the conclusions are given in Sect. 6.

2. Literature Review

A comprehensive review of existing methods in multiperson human behavior recognition (HBR) is essential to contextualize our group-centric pipeline, which aggregates visual cues for understanding collective sports behavior. Recent advances in sensing technologies have diversified approaches across RGB/multiview systems emphasizing appearance and spatial fusion, event-based methods prioritizing temporal precision via neuromorphic sensors, and radar and millimeter-wave (mmWave) sensing techniques leveraging nonvisual motion

signatures for privacy-preserving analysis. These paradigms address key challenges such as occlusions and dynamics in group settings but often overlook scalable aggregation for untrimmed videos. Table 1 shows the key papers in each domain, highlighting contributions, datasets, performance, strengths, and limitations, providing a comparative foundation for our proposed RGB-based framework.

Table 1
Related approaches to the proposed framework.

Domain	Authors	Key contributions	Dataset	Strengths	Limitations
RGB/Multiview	Cioppa *et al.*[5]	Expanded dataset with 1.3M spatial annotations for multiview soccer analysis, including player re-ID and team affiliations	SoccerNet-v3 (33986 soccer images)	Enables reproducible multiview tasks such as localization and team analysis	Limited to soccer; requires multicamera calibration
	Komorowski and Kurzejamski[6]	Deep method aggregating heatmaps to bird's-eye view; LSTM-GNN for dynamics and interactions in long-shots	Custom long-shot soccer videos	Robust to occlusions via spatial fusion; no preprocessing needed	Relies on calibrated cameras; synthetic data dependence
	Scott *et al.*[7]	Largest multisport multicamera dataset for full-pitch tracking, with 4.3M bounding boxes	TeamTrack (279900 frames: soccer, basketball, handball)	Comprehensive benchmarking for occlusions and group movements	Sports specific; annotation challenges with similar appearances
Event-based	Gao *et al.*[8]	CeleX-HAR dataset (124K sequences, 150 actions); EVMamba framework for spatio-temporal scanning on events	CeleX-HAR (event streams)	Low energy, motion-robust for low light; 20+ baselines provided	Single-person focus; sparse spatial resolution
	Adra *et al.*[9]	Survey on event cameras for behavior/facial analysis, covering representations, datasets, and applications	Various event datasets/ simulators	Highlights privacy/efficiency; future directions for hybrid fusion	Lacks empirical multi-person evaluation; noise compatibility issues
	Wang *et al.*[10]	HyperMV framework with hypergraph NN for multiview event fusion; THUMV-EACT-50 dataset release	THUMV-EACT-50 (50 actions, 6 views)	Addresses semantic misalignment; strong generalization	Hypergraph complexity for real-time; multiview dependence
Radar/mmWave	Zeng *et al.*[11]	Group tracking algorithm with extended Kalman filter; 3D-CNN-LSTM for point cloud classification	Custom 5-activity scenes (up to 3 persons)	Privacy-safe real-time deployment; handles sparsity	Limited to small groups; Kalman sensitivity to overlaps
	Wu *et al.*[12]	Feature mapping to time-domain maps; SE-DRAE-CNN for classification	Indoor/aquatic datasets	Robust to distance/ headcount; sparsity mitigation via stacking	Assumes quasi-static scenes; non-line-of-sight limitations
	Dang *et al.*[13]	Hybrid filtering + DBSCAN/ Hungarian tracking; CNN-LSTM for features	Custom gesturing activities	Low cost, robust to radar positions/ distances	Clustering sensitivity to density; small-group focus

3. Proposed Framework

RGB cameras act as low-cost visual sensors for capturing complex human activities in sensor-based systems. The proposed framework as displayed in Fig. 1 converts long, high-volume video streams into compact, discriminative representations using motion-aware sampling and hybrid feature modeling, enabling scalable processing under computational and storage constraints. Multiperson sports are classified from MultiSports RGB frames by aggregating collective dynamics, such as synchronization and pose diversity, without per-action labels. Preprocessing applies denoising and normalization, followed by motion-aware keyframe selection. Silhouettes are detected, segmented, tracked, and represented with skeletons. Motion, edge, deformation, and pose features are fused via attention, optimized through contrastive learning, and hierarchically classified to emphasize emergent group behavior.

3.1 Preprocessing

In the preprocessing phase for MultiSports videos, raw RGB frames are enhanced using total variation (TV)-based denoising to suppress sensor noise and minor illumination fluctuations while preserving important structural edges. This step improves the visual quality of the frames and increases the robustness of subsequent silhouette extraction and feature computation. The denoising is implemented using OpenCV's efficient optimization-based algorithm.

After denoising, the frame intensities are normalized using min–max scaling to map pixel values into a fixed range of [0, 1]. This normalization ensures consistent input scaling across all videos and stabilizes the learning behavior of the convolutional neural network used for deep feature extraction.

To handle lengthy clips (avg. 182.5 s at 25 fps, ~4500 frames), a motion-aware keyframe selection extracts $N_k = 30$ keyframes from T total frames. The video is divided into $N_s = 20$ equal segments S_i. Motion saliency per frame t is

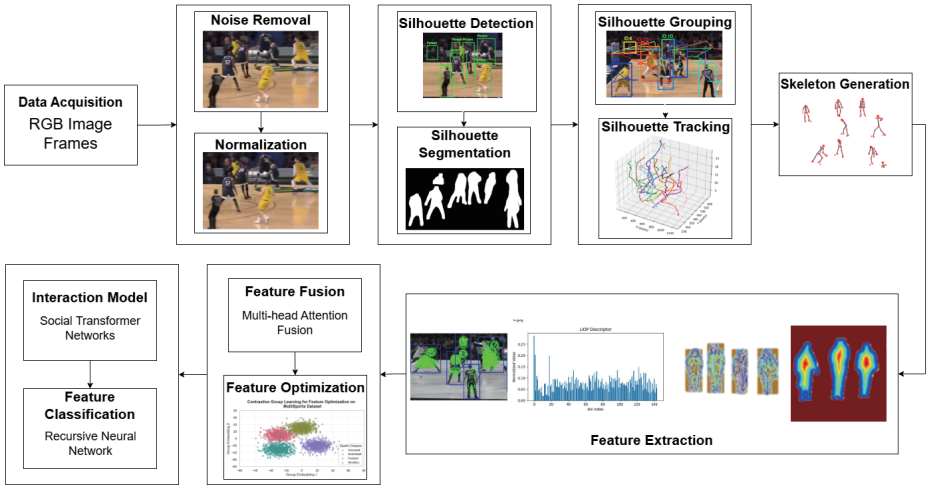


Fig. 1. (Color online) Architecture flow of our proposed system over MultiSports dataset.

$$m_t = \sum_p |G_t(p) - G_{t-1}(p)|, \quad t = 2, \ldots, T, \tag{1}$$

where G_t is the grayscale downsampled frame. The number of temporal segments was set to N_s = 20 to ensure a uniform temporal coverage of long-duration videos while maintaining sufficient granularity to capture variations in motion intensity. Empirically, this segmentation prevents keyframe concentration in short high-motion intervals and avoids excessive fragmentation that can lead to unstable motion estimates. This choice provides a practical balance between temporal resolution, robustness, and computational efficiency for motion-aware keyframe selection. The allocation of a_i to S_i is proportional:

$$a_i = \text{round}\left(\frac{\sum_{t \in S_i} m_t}{\sum_{t=1}^{T} m_t} \cdot N_k \right), \tag{2}$$

adjusted to sum exactly N_k.

Per segment, ResNet-50 features $F_i \in \mathbb{R}^{L \times 2048}$, where $L = |S_i|$, and undergoes K-means clustering ($k = a_i$):

$$C = \arg\min_C \sum_j \min_l \| f_j - c_l \|_2^2 . \tag{3}$$

Original frames nearest to the centers are selected and sorted temporally. Validation shows ~90% frame reduction with >95% motion variance retention, boosting efficiency for downstream analysis.

3.2 Multiperson silhouette and pose modeling

3.2.1 Silhouette detection

Silhouette detection identifies bounding boxes around human instances in each keyframe using Faster R-CNN, a two-stage detector that combines a region proposal network (RPN) with a classification/regression head for precise localization. Given an input frame $I \in \mathbb{R}^{H \times W \times 3}$, the backbone (ResNet-101) extracts feature maps $F = \text{Conv}(I) \in \mathbb{R}^{H' \times W' \times C}$. The RPN generates candidate regions by sliding a 3 × 3 window over F, predicting objectness scores P_o and box refinements Δb via sibling fully connected layers:

$$p_o = \sigma(W_o \cdot feat + b_o), \Delta b = W_b \cdot feat + b_b, \tag{4}$$

where σ is the sigmoid function, yielding ~2000 proposals per image after nonmaximum suppression (NMS) at IoU threshold 0.7. These proposals are pooled (via RoIAlign) into fixed-

size 7 × 7 features and fed to the detection head, which outputs class probabilities p_c (human vs background) and refined boxes $b = (x, y, w, h)$ with softmax-normalized confidence.

$$p_c = \text{softmax}\left(W_c \cdot \text{RoI}\left(F\right) + b_c\right) \tag{5}$$

We fine-tune using common objects in contex (COCO) dataset for human class, with thresholding detections at 0.5 confidence, yielding ~4−8 boxes per sports frame, which serve as initial anchors for segmentation and tracking. Figure 2 shows the results.

3.2.2 Silhouette segmentation

Building on detected boxes, silhouette segmentation refines human regions into pixel-level masks using Mask R-CNN, an extension of Faster R-CNN that adds a mask prediction branch for instance segmentation. For each RoI from detection, the model outputs a binary mask $M \in \{0,1\}^{28\times28}$ alongside the box and class, via a separate FCN head on RoI features RoI(F):

$$M = \sigma\left(\text{Deconv}\left(\text{RoI}\left(F\right)\right)\right), \tag{6}$$

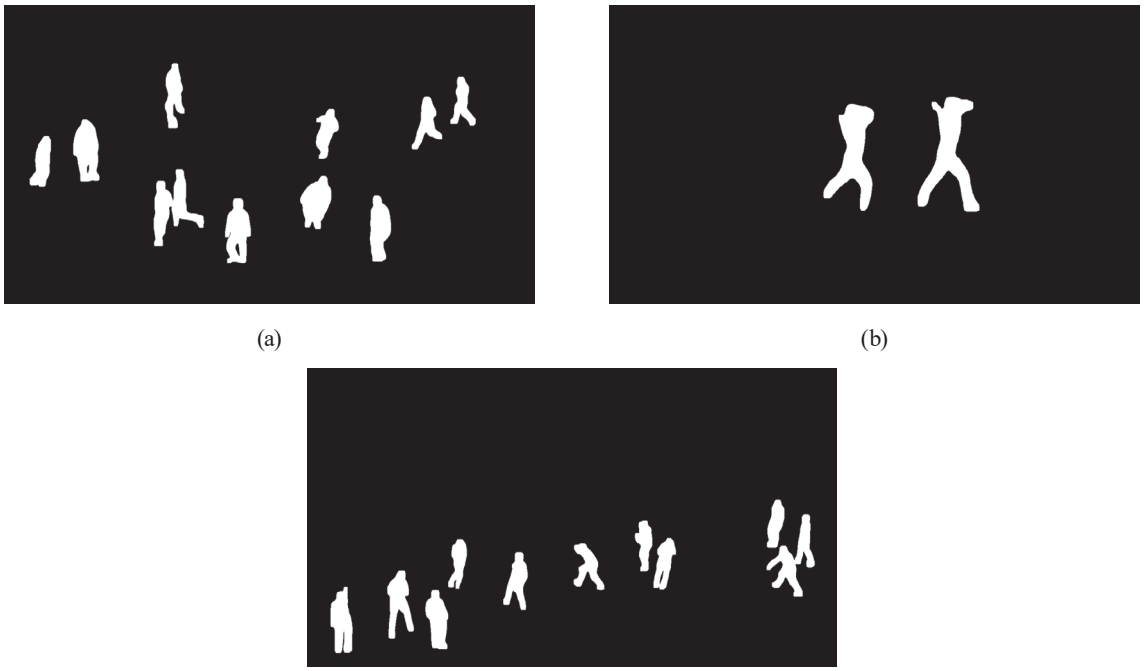
where the deconvolutional layers up sample to mask resolution. Figure 3 shows the results. The branch is trained with binary cross-entropy loss over the mask's 784 pixels ($N = 28 \times 28$):

$$\mathcal{L}_m = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log \hat{y}_n + \left(1-y_n\right)\log\left(1-\hat{y}_n\right)\right], \tag{7}$$

where y_n and $\hat{y}_n$ are ground-truth and predicted pixel labels. Masks are resized and aligned to the original box, yielding precise silhouette contours that delineate body shapes, excluding backgrounds and overlapping players. This enables accurate per-person feature extraction, with average mask IoU > 0.75 on MultiSports validation, mitigating partial occlusions in team sports.

(a) (b)

Fig. 2. (Color online) Silhouette detection results: (a) aerobics and (b) football over MultiSports dataset.



(a) (b)

(c)

Fig. 3. Silhouette segmentation results: (a) football, (b) aerobics, and (c) volleyball over MultiSports dataset.

3.2.3 Silhouette tracking

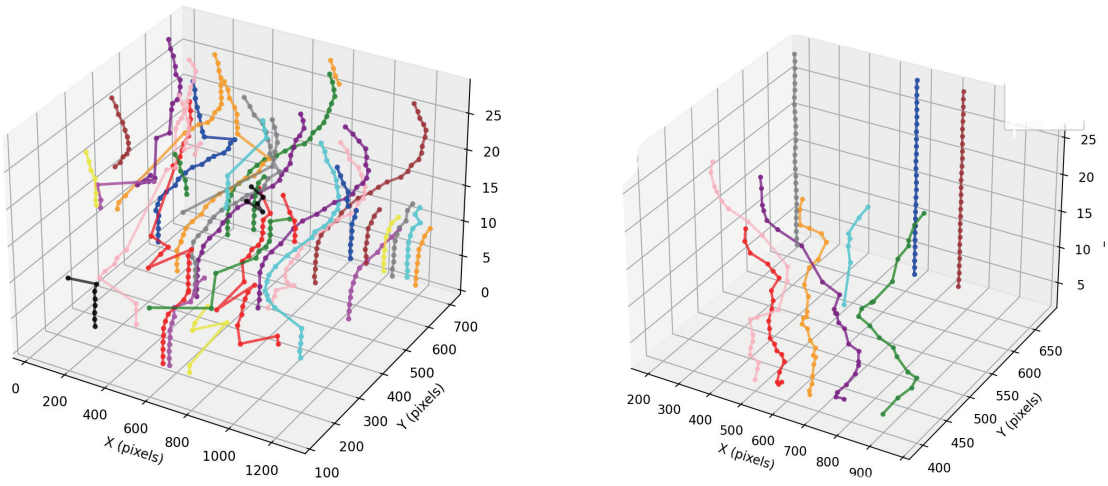
To maintain identity consistency across keyframes, we track silhouettes using a hybrid approach combining the Hungarian algorithm for global assignment and centroid distance for affinity computation. Let $\mathcal{D}_t = \{d_i^t\}_{i=1}^{K_t}$ be detections (boxes + masks) at keyframe t, and $\mathcal{T}_{t-1} = \{tr_j^{t-1}\}_{j=1}^{K_{t-1}}$ tracks from the prior frame. Centroid positions $c_i^t = \left(x_i^t + w_i^t/2, y_i^t + h_i^t/2\right)$ are extracted from boxes, and pairwise costs are computed as Euclidean distances augmented with IoU overlap:

$$\text{cost}(i,j) = \| c_i^t - c_j^{t-1} \|_2 + \lambda\left(1 - \text{IoU}\left(d_i^t, b_j^{t-1}\right)\right), \tag{8}$$

with $\lambda = 0.5$ balancing motion and overlap. The Hungarian algorithm solves the linear assignment problem to minimize total cost,

$$\min_{\pi} \sum_{i=1}^{K_t} \text{cost}\left(i, \pi(i)\right), \pi : \{1,\ldots,K_t\} \to \{1,\ldots,K_{t-1}\} \tag{9}$$

via the Jonker–Volgenant implementation for $O(K^3)$ efficiency (K ≈ 5). Unmatched tracks are terminated (if >5 frames are missed), and new detections initialize tracks, yielding temporally coherent silhouette trajectories $\{\tau_k = \left(d_k^1, \ldots, d_k^{N_k}\right)\}_{k=1}^{K}$ per video, with multiple object tracking accuracy (*MOTA*) > 0.85 on validation sequences. Figure 4 shows the results.



(a) (b)

Fig. 4. (Color online) Silhouette tracking results: (a) football and (b) aerobics over MultiSports dataset.

3.2.4 Skeleton generation

Skeleton generation produces 2D pose estimates for tracked silhouettes using Multi-person AlphaPose, a bottom-up approach that detects keypoints via stacked hourglass networks and groups them to persons via pose-guided affinity. For each frame, keypoints $P = \{p_m\}_{m=1}^{17}$ (COCO format: nose, shoulders, etc.) are detected with heatmap regression:

$$H_m = soft\max\left(UGN\left(F\right)\right), p_m = \arg\max H_m, \tag{10}$$

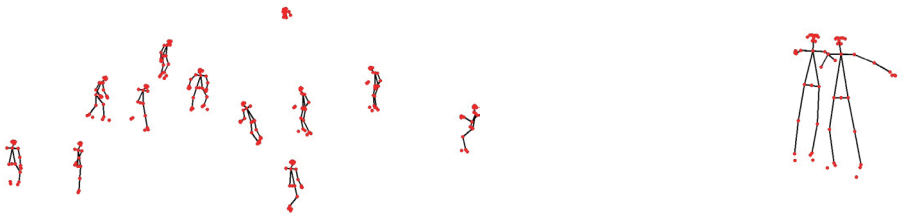
where UGN denotes the cascaded hourglass. Part affinity fields (PAFs) encode limb associations, solved via integer linear programming for grouping:

$$\max \sum_e w_e \cdot \delta_e, \text{ s.t. no overlapping limbs,} \tag{11}$$

with edge weights w_e from PAFs. Poses are assigned to tracks via nearest-centroid matching, yielding per-track skeleton sequences $\{S_k = \left(P_k^1, \ldots, P_k^{N_k}\right)\}_{k=1}^{K}$, capturing joint dynamics for interaction modeling. Figure 5 illustrates the results.

3.3 Feature extraction

The feature extraction module derives handcrafted spatiotemporal, local appearance, shape deformation, and global contour descriptors from per-person silhouettes and skeletons across keyframes, initially per tracked individual for precision, then aggregated into low-dimensional sequence descriptors across all persons and frames.



(a) (b)

Fig. 5. (Color online) Skeleton generation results: (a) football and (b) aerobics over MultiSports dataset.

3.3.1 Dense trajectories

Dense trajectories capture collective motion patterns by densely sampling and tracking points within per-person silhouette regions across keyframes, with histograms and statistical measures aggregated into sequence-level descriptors that characterize activity flow, directional biases, and temporal variability distinguishing different sports via unique motion rhythms (e.g., erratic basketball flows versus unidirectional soccer persistence).

For individual k at keyframe t, points $x_t = (x_t, y_t)$ are sampled at 5-pixel intervals inside M_t^k, near silhouette boundaries to emphasize limb and torso motion. Tracking forward for $L = 15$ frames via Farnebäck optical flow $\mathrm{u}_t = (u_t, v_t)$ (quadratic fitting for sub-pixel accuracy) yields trajectories $T_i = \{\mathrm{x}_t, \mathrm{u}_t\}_{t=1}^{L}$. Short or erratic paths ($|T_i| < L$) are discarded to retain persistent motion patterns.

Each trajectory is encoded into a 426-dimensional descriptor using normalized histograms over a 16 × 16 spatial pyramid (2 scales, 3 subwindows):

$$d_{HoG} \in \mathbb{R}^{96}\left(8 \text{ orientation bins} \times 2 \text{ scales} \times 3 \text{ subwindows}\right), \tag{12}$$

$$d_{HoF} \in \mathbb{R}^{90}\left(9 \text{ angle bins} \times 2 \text{ scales} \times 3 \text{ subwindows on } \theta_t = \operatorname{atan2}\left(v_t, u_t\right)\right), \tag{13}$$

$$d_{MBH} \in \mathbb{R}^{240}\left(8 \text{ orientation bins on } u \text{ / } v \text{ gradients to suppress global motion}\right). \tag{14}$$

The descriptors are ℓ_2-normalized and concatenated. Global pooling is then performed across all valid dense trajectories, persons, and keyframes, computing the mean and standard deviation of flow magnitudes $\mathbb{E}\left[\| \mathrm{u}_t \|_2\right]$ and $\sigma\left[\| \mathrm{u}_t \|_2\right]$, sum-pooled normalized HoG/HoF orientation histograms, and an 8-sector trajectory-endpoint direction histogram. This produces a fixed-length dense trajectory descriptor $f_{DT} \in \mathbb{R}^{500}$. Figure 6 illustrates representative results.

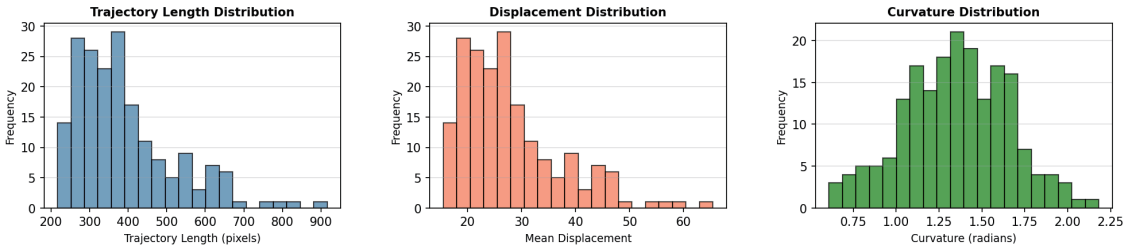


Fig. 6. (Color online) Dense trajectory results on volleyball over MultiSports dataset.

3.3.2 Oriented FAST and rotated BRIEF (ORB)

ORB keypoints identify scale- and rotation-invariant interest points on silhouette boundaries, aggregating interframe matches into bag-of-words representations augmented with density and persistence metrics to profile edge complexity and textural motifs as sport-distinctive patterns.

Within a person's mask M_t^k, FAST candidates are detected by scanning a 16-pixel-radius circle for 12 contiguous pixels exceeding intensity threshold $\tau = 20$ relative to the center, yielding $p_j = (x_j, y_j)$. Orientation ϕ_j is determined using the intensity-weighted centroid:

$$\phi_j = \operatorname{atan2}\left(\frac{\sum r_y I(r)}{\sum I(r)}, \frac{\sum r_x I(r)}{\sum I(r)}\right), \tag{15}$$

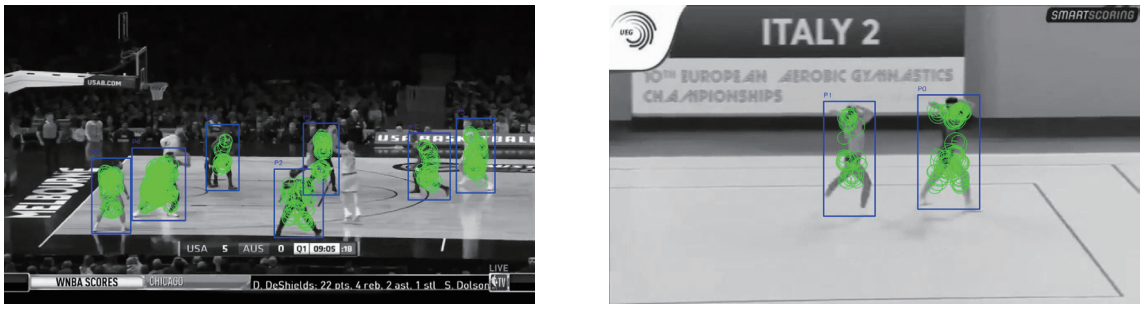
where r = (r_x, r_y) offsets in a 31-pixel patch. Descriptors are 256-bit binary strings from rotated BRIEF:

$$d(p_j) = \sum_{b=1}^{128} \mathbb{I}\left(I\left(p_j + R_{\phi_j}(o_b)\right) < I\left(p_j + R_{\phi_j}(-o_b)\right)\right) \cdot 2^{b-1}. \tag{16}$$

With the R_{ϕ_j} rotation matrix and o_b offsets, only keypoints with a Harris corner response greater than 0.01 are retained, and their corresponding 256-bit rotated BRIEF descriptors are preserved. Inter-frame Hamming-distance matching is then used to accumulate a bag-of-words histogram of retained ORB descriptors, augmented with keypoint density and temporal persistence statistics, forming the final ORB feature vector $f_{\text{ORB}} \in \mathbb{R}^{1050}$. Figure 7 shows the results.

3.3.3 Hausdorff distance features

Hausdorff distance features measure silhouette contour deformations to capture temporal motion intensity, interperson synchronization, and dynamic periodicity, aggregated from per-person pairwise comparisons to discriminate different sports by group coordination levels. Contours C_t^k are extracted per person k via Suzuki border-following on M_t^k. The frame-to-frame-directed Hausdorff distance feature for person k is computed as



(a) (b)

Fig. 7. (Color online) ORB features for (a) basketball and (b) aerobics over MultiSports dataset.

$$h\left(C_t^k, C_{t-1}^k\right) = \max_{c \in C_t^k} \min_{c' \in C_{t-1}^k} \| c - c' \|_2, \tag{17}$$

with symmetric $H_t^k = \max\left\{h\left(C_t^k, C_{t-1}^k\right), h\left(C_{t-1}^k, C_t^k\right)\right\}$ and partial variant (top-20% outliers) for robustness, normalized by contour length. Per-sequence extraction includes temporal HD statistics (mean/std/max/min/range of $\left\{H_t^k\right\}$ over persons/frames), interperson HD (mean/std/variance of pairwise H within frames), motion derivatives (mean/std of ΔH_t), and periodicity (dominant frequency/strength via FFT/autocorrelation), plus peak metrics (count/mean intervals of high-H events). This forms $f_{\mathrm{HD}} \in \mathbb{R}^{20}$, emphasizing synchronized highs in aerobics versus variable independents in basketball. Figure 8 shows the results.

3.3.4 Zernike moments

As shown in Fig. 9, Zernike moments provide rotation-invariant global shape descriptors for silhouettes, extracted per person and aggregated to sequence-level statistics that profile the sport's pose vocabulary diversity, temporal evolution, and interperson similarity without relying on action labels. For centered, unit-disk-normalized binary mask M_t^k, moments A_n^m (order $n = 0$ to 8, repetition $m = -n$ to n step 2) are projected via orthogonal polynomials:

$$A_n^m = \frac{n+1}{\pi} \iint_{M_t^k} Z_n^m(r,\theta) dr d\theta, \tag{18}$$

with radial

$$R_n^m(r) = \sum_{s=0}^{(n-|m|)/2} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} r^{n-2s} \tag{19}$$

and angular $\cos(m\theta)$ or $\sin(|m|\theta)$. Magnitudes $\left|A_n^m\right|$ ensure invariance (36 coefficients). Per-frame/person extraction yields vectors, aggregated sequence-wide as mean/std moments

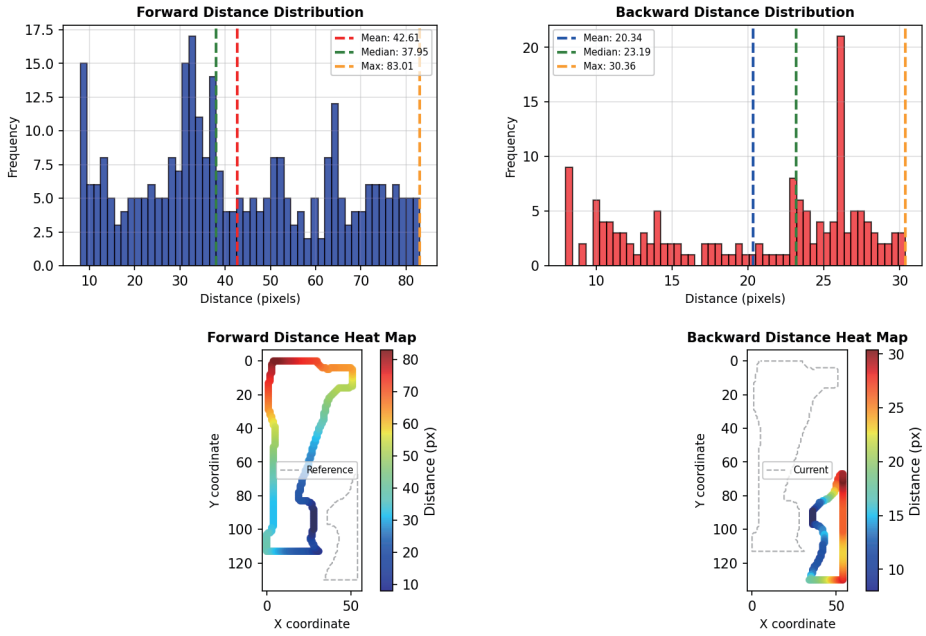


Fig. 8. (Color online) Hausdorff distance features for basketball over MultiSports dataset.

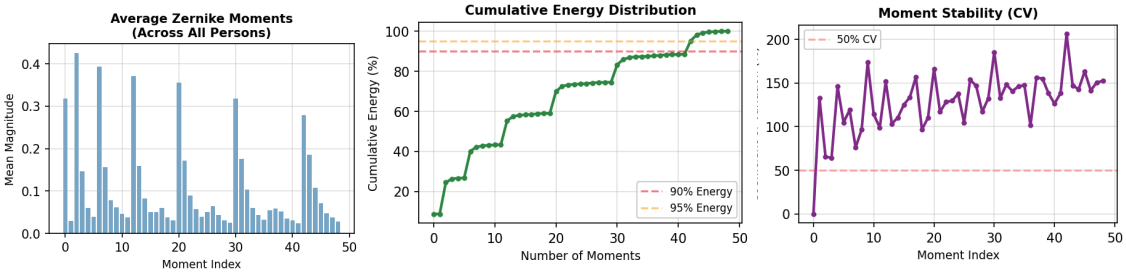


Fig. 9. (Color online) Zernike moments features for volleyball over MultiSports dataset.

($\mathbb{E}\left[\left|A_n^m\right|\right]$, $\sigma\left[\left|A_n^m\right|\right]$), temporal derivatives (mean/std of $\Delta\left|A_n^m\right|$), pose diversity, and interperson similarity. The resulting descriptor is $f_{\mathrm{ZM}} \in \mathbb{R}^{100}$, enabling discrimination between high-diversity pose patterns in aerobics and low-variability running-dominated patterns in football.

3.4 Feature fusion: Multihead attention fusion

Multihead attention fusion employs a transformer-inspired mechanism to dynamically fuse the heterogeneous feature sets $F = [f_{DT}, f_{ORB}, f_{HD}, f_{ZM}] \in \mathbb{R}^{4\times D}$ (with D being the concatenated dimension), treating each as a sequence token and attending across them to produce a context-aware global vector. The input is projected into query $Q = FW_Q$, key $K = FW_K$, and value $V = FW_V$ matrices (via learned $W \in \mathbb{R}^{D\times d_k}$, $d_k = D/h$, for $h = 4$ heads), with scaled dot-product attention per head,

$$\text{Attention}\left(Q_h, K_h, V_h\right) = \text{softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}}\right) V_h, \tag{20}$$

concatenated across heads and projected as $O = \text{Concat}\left(head_1, \ldots, head_h\right) W_O$ ($W_O \in \mathbb{R}^{hd_k \times D}$). Positional encodings $PE \in \mathbb{R}^{4 \times D}$ (sine-based for modality order) are added to F pre-attention, enabling the model to emphasize motion-dominant features (e.g., DT/HD for football) over static ones. The output $o \in \mathbb{R}^D$ is L2-normalized and fed to a two-layer position-wise feed-forward network consisting of two linear layers with a ReLU activation and residual connections for refinement:

$$\text{ffn}\left(\text{o}\right) = \max\left(0, \text{o}W_1 + b_1\right) W_2 + b_2, \tag{21}$$

with dropout (0.1) and layer norm. Trained end-to-end with cross-entropy on sport labels, this yields a fused $\text{f}_{\text{fused}} \in \mathbb{R}^D$ per video, improving classification by 5–8% over simple concatenation.

3.5 Feature optimization: Contrastive learning

The feature optimization module uses contrastive learning on unlabeled MultiSports interactions to refine fused representations, enhancing intrasport compactness and intersport separation for the robust, label-efficient classification of group dynamics (e.g., team coordination).

Contrastive group learning optimizes the fused features f_{fused} by projecting them into a low-dimensional space where group instances from the same sport are clustered closely together, while those from different sports are separated, formulated as a noise-contrastive estimation task over positive (same-sport) and negative (different-sport) pairs. Each video's feature is treated as a group anchor $\text{a}_i \in \mathbb{R}^D$, augmented with two views a_i^+ and a_i^- (via random dropout and jittering on temporal statistics), projected through a two-layer MLP encoder $g(\cdot): \mathbb{R}^D \to \mathbb{R}^d$ (d = 128) with ReLU and L2-normalization to yield embeddings $z_i = g(a_i)$, $\text{z}_i^+ = g\left(a_i^+\right)$, and negatives $\{\text{z}_j^-\}_{j \neq i}$ sampled from the batch (e.g., 256 videos, 4 negatives each).

The objective minimizes the noise-contrastive estimation (NCE) loss, encouraging alignment between positive pairs and uniformity across negatives:

$$\mathcal{L}_i = -\log \frac{\exp\left(\text{z}_i \cdot \text{z}_i^+ / \tau\right)}{\exp\left(\text{z}_i \cdot \text{z}_i^+ / \tau\right) + \sum_{j=1}^{N} exp\left(z_i \cdot z_j^- / \tau\right)}, \tag{22}$$

where τ = 0.07 is the temperature scaling alignment strength and N = 4 negatives. The total loss is averaged over the batch:

$$\mathcal{L} = \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_i, \tag{23}$$

with $B = 64$ batch size. Group-level positives are dynamically sampled by masking intravideo persons (e.g., shuffling trajectories to simulate occlusions), ensuring the learned space captures emergent sport signatures such as synchronization variance. Optimization is performed with AdamW ($\eta = 1e^{-4}$, weight decay $1e^{-2}$) for 100 epochs, yielding optimized $f_{opt} = g(f_{fused}) \in \mathbb{R}^{128}$, which boosts downstream classification accuracy by 3–5% on validation by reducing feature overlap across sports. Figure 10 shows the results.

3.6 Interaction modeling: Social transformer networks

Social transformer networks model interactions via a graph transformer architecture that processes per-frame person embeddings as nodes, attending over spatiotemporal relations to infer group semantics for sport classification. For each keyframe t, tracked persons yield node features $n_k^t = \left[f_{opt}^k, p_k^t \right] \in \mathbb{R}^{D+17}$ (fused features plus COCO keypoints p_k^t), forming a set $\mathcal{N}^t = \{n_k^t\}_{k=1}^K$ ($K \approx 5$). Edge affinities are computed implicitly via multihead self-attention, with positional encodings incorporating relative distances $d_{ij}^t = \| c_i^t - c_j^t \|_2$ (centroids) and joint angles for social cues.

The transformer layers (4 layers, 8 heads) apply scaled dot-product attention on projected queries/keys/values:

$$\mathrm{Attention}\left(Q,K,V\right) = \mathrm{softmax}\left(\frac{QK^T + E}{\sqrt{d_k}}\right)V, \tag{24}$$

where $\mathrm{E} \in \mathbb{R}^{K \times K}$ is an edge bias matrix encoding $e_{ij} = -d_{ij}^t / \sigma_d + \cos\left(\alpha_{ij}^t\right)$ ($\sigma_d = 50$, α_{ij}^t: average joint angle difference), promoting attention on nearby/interacting pairs. Outputs are pooled via mean over nodes to $i^t \in \mathbb{R}^D$, then fed to a temporal GRU (*hidden* = 256) across $N_k = 30$ keyframes:

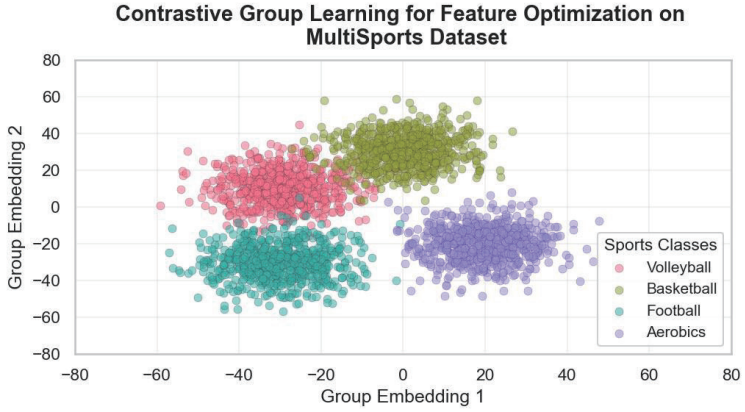


Fig. 10. (Color online) Feature optimization over MultiSports dataset.

$$h_t = \mathrm{GRU}\left(\left[i^t, h_{t-1} \right]\right), \tag{25}$$

yielding the final interaction embedding $f_{int} = h_{N_k} \in \mathbb{R}^{256}$. Trained with auxiliary triplet loss on interaction triples (anchor video, positive same-sport, negative different), this enriches f_{opt} for classification, improving group-aware accuracy by 4–6% on validation.

3.7 Feature classification: RecNN

The RecNN classifies the input embedding by recursively composing representations over a binary tree structure derived from the temporal sequence of keyframes, where leaves are per-frame features and internal nodes aggregate child states to model hierarchical dependences in group behaviors. The tree is constructed by the balanced binary partitioning of the N_k = 30 keyframes, yielding depth $\log_2 N_k \approx 5$ and $2N_k - 1$ nodes. At each leaf node l (frame t), the initial vector is $v_l = f_{cls}^t \in \mathbb{R}^{384}$, augmented with parent direction embeddings $\mathrm{p}_l \in \mathbb{R}^{10}$ (one-hot for left/right). For an internal node n with children c_1, c_2, the representation v_n is computed recursively via a composition function:

$$v_n = \tanh\left(W \cdot \left[v_{c_1}; v_{c_2}; p_n \right] + b\right), \tag{26}$$

where $W \in \mathbb{R}^{256\times(384+384+10)}$, $b \in \mathbb{R}^{256}$ are learned parameters, and [;] denotes concatenation, reduced to 256 dimensions per node for efficiency. This bottom-up recursion propagates until the root $v_{root} \in \mathbb{R}^{256}$, which encodes a global sequence of semantics.

Classification at the root applies a linear layer followed by SoftMax:

$$\hat{y} = \mathrm{softmax}\left(V v_{root} + b_y\right), \quad V \in \mathbb{R}^{4\times256}, \tag{27}$$

with cross-entropy loss $\mathcal{L} = -\sum y \log \hat{y}$ over one-hot labels y. Trained with Adam ($\eta = 1e^{-3}$) for 50 epochs and dropout (0.2), the RecNN achieves 92.3% best accuracy.

4. Results and Discussion

The pipeline was evaluated on the MultiSports dataset using 10-fold cross-validation across football, volleyball, basketball, and aerobics, implemented in PyCharm on an NVIDIA RTX A2000 GPU. Performance was assessed using detection (mAP), tracking (*MOTA*, *HOTA*, *IDF*1, *AssA*, *DetA*), and classification (accuracy, precision, recall, F-1 score, AUC-ROC) metrics. Faster R-CNN achieved 78.5% mAP@0.5, while spatiotemporal feature tubelets outperformed MultiSports baselines by 5–7% (Table 2). Hungarian-based tracking maintained identity consistency with *MOTA* = 79.4%, *HOTA* = 76.1%, and *IDF*1 = 78.2% (Table 3), and the system ran at 36.5 FPS, enabling real-time group sports analysis.

Table 2
Comparative performance analysis of state-of-the-art action detection methods and our proposed group-aware approach over MultiSports dataset.

Method	mAP@0.5	f-mAP@0.5	v-mAP@0.2	v-mAP@0.5	v-mAP@[0.1:0.9]
YOWO[2,14]	–	25.2	12.9	9.7	–
SlowFast-R101 + PCCA[15]	–	42.2	41.0	20.0	20.9
MultiSport Baseline[14]	–	49.6	54.1	31.3	28.9
MultiSport Baseline + Tracks[14]	–	50.6	56.3	33.0	30.9
TAAD + TCN[14]	–	55.3	60.6	37.0	33.7
Proposed Method (Ours)	78.5	57.1	62.4	38.9	35.1

Table 3
Evaluation of tracking consistency and efficiency for MultiSports dataset.

Metric	Value (%)	Interpretation
HOTA	76.1 ± 1.4	Stable overall tracking accuracy; low variance across folds indicates consistent spatiotemporal matching
MOTA	79.4 ± 1.7	High multi-object tracking accuracy with minimal ID switches and false alarms
IDF1	78.2 ± 1.3	Consistent identity preservation across players and sports sequences
AssA	76.8 ± 1.5	Robust temporal association, confirming reliable frame-to-frame continuity
DetA	78.6 ± 1.6	Strong detection contribution to tracking, driven by refined Mask R-CNN segmentation
FPS	36.5 ± 0.8	Near real-time inference efficiency with minor variation due to scene complexity

The RecNN-based classifier achieved 91.25% overall accuracy across four sports classes, effectively distinguishing collective behaviors (Table 4). Table 5 shows precision, recall, and F1-score. ROC analysis (Fig. 11) confirms high discriminability, with per-class and macro-AUC values of 0.94, indicating high true-positive rates and low false positives.

To empirically justify integrating handcrafted features with deep representations, an ablation study was conducted on the MultiSports dataset (Table 6). The deep-only model (DM), which excludes handcrafted motion and shape descriptors, achieves 86.80% accuracy, showing that learned pose and appearance features alone are insufficient for complex group sports dynamics, whereas the full hybrid model (FM) attains the highest accuracy of 91.25%.

5. Discussion

Despite the promising results, this study has several limitations. First, the proposed framework is evaluated using publicly available RGB video datasets rather than custom-designed sensing systems, which may limit adaptability to unconstrained real-world environments. Second, the motion-aware keyframe selection strategy employs fixed parameter settings that, while effective, may not be optimal for all sports or activity types. Additionally, the current framework assumes relatively stable camera viewpoints and does not explicitly address severe occlusions or abrupt scene changes. Future research will focus on extending the framework to multimodal sensor data, such as depth or wearable signals, and on developing adaptive keyframe

Table 4
Confusion matrix showing per Class accuracy on MultiSports dataset.

Class	Football	Volleyball	Baseball	Aerobics
Football	92	3	2	3
Volleyball	4	91	3	2
Basketball	3	4	90	3
Aerobics	2	2	4	92
Mean		91.25		

Table 5
Class-wise precision, recall, and F1-score for the proposed framework over MultiSports dataset.

Class	Precision	Recall	F1-score
Football	0.911	0.920	0.915
Volleyball	0.901	0.910	0.905
Basketball	0.900	0.900	0.900
Aerobics	0.928	0.910	0.919
Mean	0.910	0.910	0.910

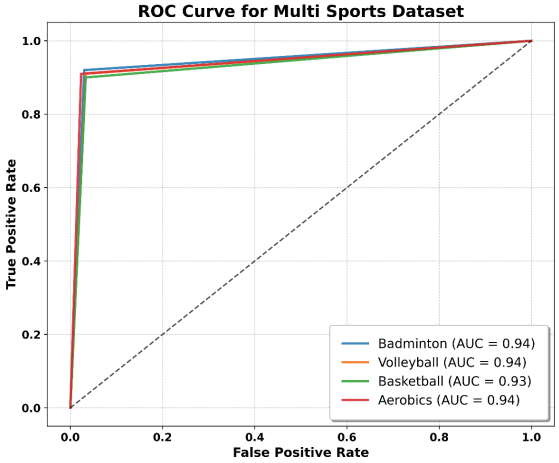


Fig. 11. (Color online) ROC curve for MultiSports dataset.

Table 6
Ablation study evaluating the contribution of deep and handcrafted feature components on the MultiSports Dataset.

Exp	PR	SD	SS	ST	SG	SK	ORB	DT	HD	ZM	FF	FO	MSD
FM	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	91.25
DM	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✓	86.80
w PR	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	91
w SD	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	89
w SS	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	87
w ST	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	90
w SG	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	91
w SK	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	86
w EG	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	90
w DT	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	89
w BR	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	91
w RG	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	90
w FF	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	89
w FO	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	89

FM = Full Model, w = without, DM = Deep-only Model, PR = Preprocessing, SD = Silhouette Detection, SS = Silhouette Segmentation, ST = Silhouette Tracking, SG = Silhouette Grouping, SK = Skeleton Keypoints Generation, ORB = Oriented FAST and rotated BRIEF, DT = Distance Transform, HD = Hausdorff distance, ZM = Zernike moments, FF = Feature Fusion, FO = Feature Optimization, MSD = MultiSports Dataset.

selection strategies that dynamically adjust to activity complexity. Further investigation into cross-dataset generalization and real-time deployment scenarios will also be pursued.

6. Conclusions

We presented a pipeline for video-level sport classification in multiperson scenarios using noncontact RGB visual sensors to aggregate per-person representations into sequence-level descriptors that capture emergent group dynamics on the MultiSports dataset. Through motion-aware keyframe selection, silhouette-based human representation, hybrid handcrafted feature aggregation, attention-based fusion, interaction modeling, and RecNN classification, the approach shifts focus from isolated actions to collective signatures such as pose diversity and synchronization, enabling robust discrimination across multiple sports. From a sensors perspective, the framework demonstrates how standard RGB vision sensors can be efficiently exploited for group behavior recognition by reducing redundancy in long video streams while preserving critical motion and interaction cues. This supports practical deployment in resource-constrained sensor-based systems, including intelligent sports facilities, surveillance, and smart environments.

Acknowledgments

This work was supported by NUIST Talent Start-up Fund (No. 1513142501062) and Jiangsu Distinguished Fund (No. R2025T07). The publication was also supported by the Open Access Initiative of the University of Bremen and the DFG via SuUB Bremen. This research was supported and funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number PNURSP2026R410, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

1 C. Feichtenhofer, H. Fan, J. Malik, and K. He: Proc. 2019 IEEE/CVF Int. Conf. Computer Vision (ICCV) (IEEE, 2019) 6201–6210. https://doi.org/10.1109/ICCV.2019.00630
2 Y. Li, Z. Wang, L. Wang, and G. Wu: Computer Vision – ECCV 2020, Lect. Notes Comput. Sci. 12361 (Springer, Cham, 2020). https://doi.org/10.1007/978-3-030-58517-4_5
3 O. Köpüklü, X. Wei, and G. Rigoll: arXiv:1911.06644 (2019). https://arxiv.org/abs/1911.06644 (accessed October 2025).
4 H. Yin, R. O. Sinnott, and G. T. Jayaputera: Artif. Intell. Rev. **57** (2024) 293. https://doi.org/10.1007/s10462-024-10934-9
5 A. Cioppa, A. Deliège, S. Giancola, B. Ghanem, and M. Van Droogenbroeck: Sci. Data **9** (2022) 1. https://doi.org/10.1038/s41597-022-01469-1 (accessed October 2025).
6 J. Komorowski and G. Kurzejamski: Proc. 2022 Int. Joint Conf. Neural Networks (IJCNN) (IEEE, 2022) 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892562
7 A. Scott, I. Uchida, N. Ding, R. Umemoto, R. Bunker, R. Kobayashi, T. Koyama, M. Onishi, Y. Kameda, and K. Fujii: arXiv:submit/5550700 (2023).
8 Y. Gao, J. Lu, S. Li, Y. Li, and S. Du: IEEE Trans. Pattern Anal. Mach. Intell. **46** (2024) 6610. https://doi.org/10.1109/TPAMI.2024.3382117
9 M. Adra, S. Melcarne, and J. Dugelay: Front. Signal Process. **5** (2025) 1585242. https://doi.org/10.3389/frsip.2025.1585242

10 X. Wang, S. Wang, P. Shao, B. Jiang, L. Zhu, and Y. Tian: arXiv:2408.09764 (2024). https://arxiv.org/abs/2408.09764 (accessed October 2025).
11 X. Zeng, Y. Shi, and A. Zhou: Proc. 2022 IEEE 8th Int. Conf. Computer Commun. (ICCC) (IEEE, 2022) 1789–1793. https://doi.org/10.1109/ICCC56324.2022.10065810
12 Z. Wu, Z. Cao, X. Yu, J. Zhu, C. Song, and Z. Xu: IEEE Sensors J. **23** (2023) 19509-19523. https://doi.org/10.1109/JSEN.2023.3283778
13 X. Dang, K. Fan, F. Li, Y. Tang, Y. Gao, and Y. Wang: Appl. Sci. **14** (2023) 7253. https://doi.org/10.3390/app14167253
14 Y. Li, L. Chen, R. He, Z. Wang, G. Wu, and L. Wang: Proc. 2021 IEEE/CVF Int. Conf. Computer Vision (ICCV) (IEEE, 2021) 13516–13525. https://doi.org/10.1109/ICCV48922.2021.01328
15 Z. Ning, Q. Xie, W. Zhou, L. Wang, and H. Li: Tech. Rep., Huawei Noah's Ark Lab and Univ. of Sci. and Technol. of China (2021). https://deeperaction.github.io/iccv21/report2/Person_Context_Cross_Attention_for_Spatio_Temporal_Action_Detection.pdf