Bachelor Thesis

Omid Naeej Nejad
610301189

1 Datasets Have Approaching, Splitting and Walking-Together Actions

This section summarizes the most suitable group activity recognition (GAR) datasets for the three target actions in this project: (1) converging walk, (2) diverging walk, and (3) walking together. The selected datasets contain social interaction labels such as approaching, splitting, meeting, and moving together, which directly correspond to the desired categories.

1.1 BEHAVE Interactions Dataset

The BEHAVE dataset is one of the most widely used resources for multi-person behavior classification. It provides detailed, ground-truthed annotations for pairwise and group interactions. Relevant classes include:

- Approach (maps to converging walk)
- Meet (maps to converging walk)
- WalkTogether (maps to walking together)
- Split (maps to diverging walk)

Other available actions include InGroup, Ignore, RunTogether, Fight, Chase, and Following. BEHAVE is a strong match because its labeled actions closely mirror the target categories of this project.

1.2 CAVIAR Dataset

The CAVIAR dataset contains surveillance scenarios with multiple social behaviors. Its annotations include person trajectories and interaction events. Relevant actions include:

- People meeting (converging walk)
- Walking together (walking together)
- Splitting up (diverging walk)

CAVIAR has been frequently used for early social behavior and group-activity recognition research, making it suitable for baseline experiments.

1.3 New Collective Activity Dataset

The New Collective dataset focuses on crowd-level and small-group activities in outdoor scenes. Its group actions closely align with the target behaviors:

- Gathering (people converge into a group)
- Dismissal (people disperse from a group)
- Walking together (group motion)

These labels naturally map to the project's categories of converging, diverging, and walking-together motion patterns.

1.4 UCLA Courtyard Dataset

The UCLA Courtyard dataset contains group-level actions in a controlled courtyard environment. While it does not explicitly include converging or diverging labels, it includes:

- Walking-together (group locomotion)

Other actions (standing in line, discussing in group, sitting together, guided tour) can serve as negative samples or auxiliary behaviors but do not directly represent convergence or divergence.

1.5 M3Act Synthetic Dataset

The M3Act dataset is a synthetic multi-view, multi-group activity dataset that allows controlled generation of group behaviors. Although it does not provide explicit converging or diverging labels, it can be used to generate and annotate the following:

- Converging walk (groups moving toward each other)
- Diverging walk (groups splitting or separating)
- Walking together (multiple synchronized trajectories)

Because it is synthetic, M3Act is valuable for pretraining, augmentation, and generating large-scale labeled patterns before fine-tuning on real datasets like BEHAVE or CAVIAR.

1.6 Summary of Dataset–Action Mapping

These datasets collectively provide strong coverage of the three required actions. BEHAVE and CAVIAR are the most directly aligned with the project's target behaviors and should be prioritized for model training and evaluation.

Table 1: Summary of datasets

Dataset	#Clips	#Classes	Resolution	Notes
BEHAVE	∼400	10	640×480	Approach, Split, WalkTogether
CAVIAR	∼52	7	384×288	Meeting, Walking together, Splitting up
New Collective	∼360	6	720×576	Gathering, Dismissal, Walking together
UCLA Courtyard	103	6	1920×1080	Walking-together only
M3Act (synthetic)	1000+	Custom	Variable	User-generated converge/diverge

2 Surveyed Literature on Group Activity Recognition

This section summarizes the major surveys on group activity recognition (GAR), focusing on their taxonomies, scope, datasets, and methodological categorizations. These surveys provide structured perspectives on how group behaviors such as converging, diverging, and walking-together actions are modeled, analyzed, and benchmarked.

2.1 List of Major Surveys

The following surveys were identified as the most comprehensive and relevant for GAR:

- A Comprehensive Review of Group Activity Recognition in Videos (2021)
- Group Activity Recognition in Computer Vision: A Comprehensive Review, Challenges, and Future Perspectives (2023)
- Deep Learning-based Group Activity Recognition in Videos (2025)
- A Comprehensive Study of Group Activity Recognition Methods and Datasets (multiple compilation-style works, 2019–2024)

Each survey proposes a slightly different taxonomy, focusing on aspects such as feature-based vs. deep architectures, scene-level vs. actor-level modeling, graphical vs. attention-based models, and benchmark datasets.

2.2 Survey 1: "A Comprehensive Review of Group Activity Recognition in Videos" (2021)

This survey provides a classic, structured taxonomy that separates GAR methods into three main categories:

- **Handcrafted Feature Methods** Trajectory-based descriptors, social-force models, relative motion patterns, spatial occupancy, and handcrafted group descriptors.
- **Deep Learning-based Methods** CNN + LSTM architectures, hierarchical LSTMs (person-level then group-level), spatiotemporal models, late fusion, and early fusion strategies.

- **Graph-based Methods** Actor–relation graphs, pairwise interaction graphs, message passing, and structural reasoning over human nodes.

The survey contains an extensive dataset table covering BEHAVE, CAVIAR, Collective Activity, UCLA Courtyard, Volleyball, and sports datasets. It explicitly describes labels such as Approach, Split, WalkTogether, Meeting, Gathering, and Dismissal, which directly relate to the classes required in this project.

2.3 Survey 2: "Group Activity Recognition in Computer Vision: A Comprehensive Review, Challenges, and Future Perspectives" (2023)

This newer survey proposes a broader and more modern taxonomy. Its major contributions include:

- **Scene-level vs. Group-level vs. Person-level Modeling** Categorizes methods based on whether the model focuses on global scene cues, group structure, or individual trajectories.
- **Context Reasoning Models** Emphasizes spatial context (layout, background), social context (inter-person distances), and temporal context (group formation or dispersion).
- **Transformer-based and Relation-aware Models** Includes self-attention models, structured attention, and multi-head relation reasoning.
- **Benchmark Comparisons Across 10+ Datasets** Provides a unified comparison of dataset properties, evaluation metrics, and model performance.

This survey is particularly useful for modern architectures (Transformers, GNNs, attention) and for situating our problem of converging/diverging actions within structured "group formation/dissolution" dynamics.

2.4 Survey 3: "Deep Learning-based Group Activity Recognition in Videos" (2025)

This survey focuses exclusively on deep architectures and organizes methods according to the computational structure of the model:

- **Hierarchical Architectures** Two-level LSTMs, multi-stage aggregation, and actor-to-group reasoning.
- **Graph Neural Networks for Group Reasoning** Graph convolution, graph attention, message passing with nodes as people and edges as interactions.
- **Transformer Models** Pure attention models for multi-person dynamics, group token aggregation, and relational transformers.

- **Weakly-supervised and Semi-supervised GAR** Methods that require only video-level labels (e.g., DFWSGAR), which is useful when detailed frame-level annotations are missing.

The survey includes a detailed taxonomy of supervisory signals and compares models on the Volleyball and Collective Activity datasets.

2.5 Survey 4: "Compilation of Group Activity Recognition Methods and Datasets" (2019–2024)

This set of shorter surveys or compilations focuses on dataset properties. Their taxonomies generally categorize:

- **Data-driven categories**: indoor vs. outdoor, crowd vs. small group.
- **Annotation style**: bounding boxes, tracklets, interaction graphs.
- **Action structure**: symmetric vs. asymmetric interactions (e.g., WalkTogether vs. Approach).
- **Application domains**: surveillance, sports, social events.

These are especially useful when comparing dataset suitability for converging/diverging actions.

2.6 Comparison of Survey Taxonomies

Survey	Architecture Taxonomy	Dataset Coverage	Interaction Focus
Review (2021)	Features / Deep / Graph	Strong	Strong (Approach/Split)
Review (2023)	Scene / Group / Person / Context	Very Strong	Moderate
Review (2025)	Hierarchical / GNN / Transformer	Moderate	Weak
Dataset Compilations	None (data-focused)	Very Strong	Strong

2.7 Summary

- The 2021 review is the best for understanding early models and datasets like BEHAVE and CAVIAR.
- The 2023 review is the best if your thesis involves relational or attention-based modeling.
- The 2025 survey is ideal for deep-learning-only architectures (GNNs, Transformers).
- Dataset compilations are extremely useful for matching specific behaviors: Approach, Split, Meeting, WalkTogether.