Social Intelligence in Computer Vision: A Review of Cues, Models, and Applications

Mohmmad M. Zare'i

University of Tehran

September 27, 2025

Outline

Introduction

- **Activity Recognition (Individual)** Identifying what a person is doing. Datasets: UCF101, Kinetics
- **Group Activity Recognition (Social)** Understanding coordinated or interacting actions within groups. Datasets: Collective Activity, Volleyball Dataset
- **Why It Matters**
  - Moves beyond visual labels $\rightarrow$ requires reasoning about *roles, relationships, intentions*.
  - Foundation for social intelligence in computer vision.
  - Applications: autonomous driving, social robotics, surveillance, human–AI collaboration.

Complexity Beyond Actions

- Group activities involve **multiple people** and their interactions.
- Requires modeling of:
  - *Roles* (leader, follower, bystander)
  - *Relationships* (friend, rival, teammate)
  - *Intentions* (cooperation, competition, avoidance)
  - *Context* (sports field, meeting room, street)
- Moves beyond simple action recognition → toward **social reasoning**.

Social and Cognitive Cues

- Key cues from **social psychology** and **cognitive science**:
  - Facial expressions $\rightarrow$ emotions
  - Body language $\rightarrow$ posture, gestures
  - Gaze direction $\rightarrow$ attention, focus
  - Proxemics $\rightarrow$ distance and spatial relationships
  - Turn-taking $\rightarrow$ conversational dynamics
- These cues provide the foundation for **interpreting group interactions**.

Technical Challenges

- Multi-person detection and tracking across frames.
- Temporal reasoning: modeling interactions over time.
- Multimodal integration: vision + audio + language.
- Ambiguity and context dependence of group behaviors.
- Generalization: handling unseen group dynamics.

From Recognition to Understanding

- Traditional goal: **recognize what is happening**.
- Emerging goal: **understand why it is happening**.
  - Infer group goals, intentions, and social context.
  - Connect low-level cues with high-level reasoning.
- This leap is the essence of **social intelligence in video**.

Literature Review

- Overview of key works in social intelligence and computer vision.
- Discussion of methodologies, datasets, and findings.
- Identification of gaps and future directions.

SoGAR: Self-supervised Spatiotemporal Attention-based Social Group Activity Recognition

- **Datasets**: Volleyball Dataset, JRDB-PAR, NBA Dataset
- **Key Contributions**:
  - Introduced a self-supervised learning framework.
  - Leveraged spatiotemporal attention for improved interaction modeling.
  - Demonstrated effectiveness on multiple group activity datasets.
- **Model Architecture**:
  - Base: Vision Transformer (ViT) and TimeSformer
  - Learning: Self-supervised pretraining on large video datasets

SoGAR

- **Methodology**:
- Uses a self-supervised transformer framework (Vision Transformer backbone) for social group activity recognition.
- Key idea: generate **local** and **global** spatio-temporal views from the same video, with variation in frame rate and spatial crop size.
- A teacher-student architecture: teacher processes global view, student processes local views. The student is trained to align its features to those of the teacher.
- Two contrastive / correspondence objectives:
  - Temporal Collaborative Learning (TCL): relate views differing in temporal resolution.
  - Spatio-temporal Cooperative Learning (SCL): relate views that differ in spatial crop + temporal sampling.
- Does **not** require actor bounding boxes or individual action labels during pre-training, reducing annotation burden.
- Uses motion as supervisory signal from RGB alone; the model becomes invariant to scale, viewpoint, and motion speed.

LaIAR: Language-Model-Guided Interpretable Video Action Reasoning

- **Datasets**: Charades, CAD-120
- **Key Contributions**:
  - Proposed a framework that integrates large language models (LLMs) for interpretable action reasoning. Utilizing knowledge transfer between LLMs and video model.
  - Utilized LLMs to generate explanations and rationales for recognized actions. Inferring high-level actions from low-level changes in relationships between actors and objects.
- **Model Architecture**:
  - R-CNN and ResNet-101 as backbones for object, category, and relation detection.
  - Relations and visual features are mapped to a joint embedding space.
  - Embeddings are fed into a dynamic token transformer (DT-Former).

LaIAR

- **Methodology**:
- Introduces a dual-branch framework: a video model and a language model, trained together so the video model learns reasoning from the language model.
- Uses relationship transitions between humans/objects as cues: visual relations (appearance, bounding boxes, spatial configuration) and semantic relations (human/object categories + relationships) are encoded.
- Both visual and semantic relations are encoded via Faster R-CNN + ResNet-101 for detecting entities, extracting features, forming human-object pairs.
- Core architecture: DT-Former (Dynamic Token Transformer) which applies adaptive token selection (spatio-temporal tokens) and then transformer layers to model relation transitions. Tokens with low importance are discarded via a Gumbel-Softmax mechanism.
- Learning scheme includes:
  - Joint visual-semantic embedding: aligning visual relation features and semantic relation features into a common space