CROWD SIMULATION WITH GUIDED REINFORCEMENT LEARNING

Nguyễn Ngọc Thạch

Trường Đại học Công nghệ Thông tin - Đại học Quốc gia TP.HCM

What ?

- We presents a reinforcement learning-based approach for simulating crowd behaviors.
- Training agents to move naturally based on real pedestrian trajectories from video input.

Why ?

- Human behavior in dense crowds is difficult to hard-code.
- Rule-based systems often fail to produce natural, flexible movement.
- Guided RL enables agents to learn movement patterns by imitating real-world behavior instead of relying on handcrafted rules.

Overview

Trajectory extraction
Guided Reinforcement Learning
Trained Policy

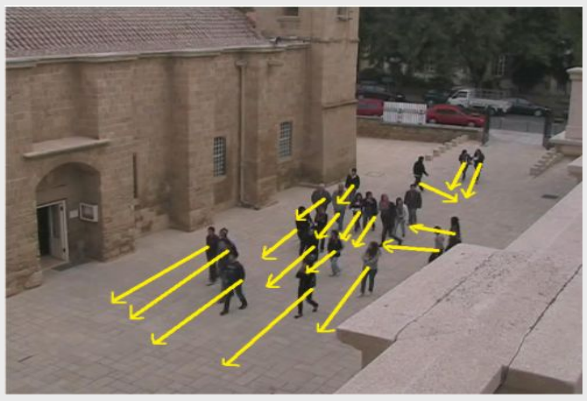

Figure 1. Yellow arrows indicating movement direction and trajectory path

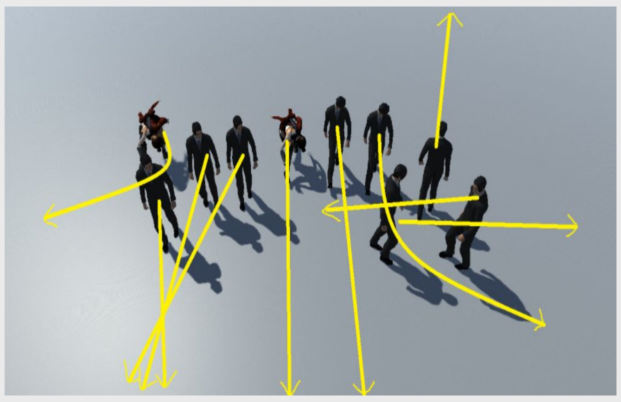

Figure 2. Mid-training scene, agents learning based on trajectory-guided rewards

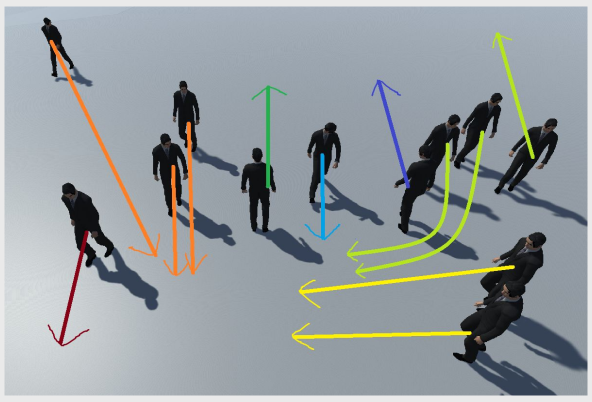

Figure 3. Post-training, agents join/split group, avoid others, and move more natural

Description

1. Trajectory Extraction

- Crowd videos are processed using YOLO and DeepSORT to detect and track pedestrians across time.
- Each person's movement is captured as a sequence of (x, y) positions over time, forming a trajectory
- These extracted trajectories serve as guidance signals for agent learning.

2. Guided Reinforcement Learning

- Agents are trained using PPO (Proximal Policy Optimization), where the reward function is shaped based on how closely the agent's behavior matches real trajectories.
- This avoids manually crafting rules and instead encourages agents to move naturally like real people.

3. Training Environment

- The environment is built in Unity using ML-Agents. Each episode simulates multiple agents initialized with goals and start positions. Agents observe their surroundings and make continuous decisions about movement.

4. Trained Policy and Generalization

- After training, agents exhibit emergent behaviors such as maintaining group cohesion, collision avoidance, and adaptive turning.
- The model is evaluated using ADE (Average Displacement Error) and FDE (Final Displacement Error).

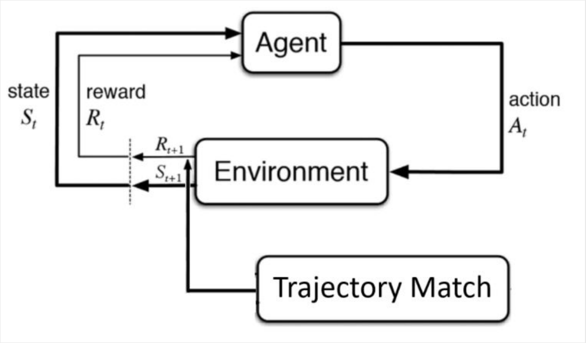

Figure 4. Standard RL loop with trajectory-based reward shaping.

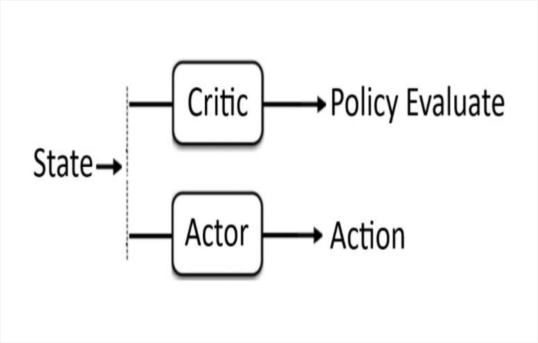

Figure 5. PPO structure with Actor–Critic design. The Actor generates actions, the Critic evaluates the value of the current policy.

NII

Nguyễn Ngọc Thạch – Trường Đại học Công Nghệ Thông Tin
TEL : 0828123416 Email : thachnn.19@grad.uit.edu.vn