Context Aware Group Activity Recognition

Avijit Dasgupta[1]

C. V. Jawahar[1]

Karteek Alahari[2]

[1] CVIT, IIIT Hyderabad, India

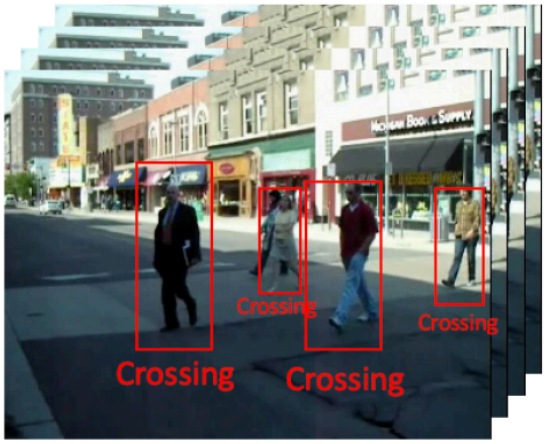
[2] THOTH, Inria, France

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Problem Definition

Given a multi-person video, the task is to infer

- actions being performed by the individuals
- their group activities



Group Activity: Crossing

Motivation

- Existing approaches rely on ***appearance*** only features
- Unable to differentiate between visually similar activities
- **Context** gives cues for group activity understanding

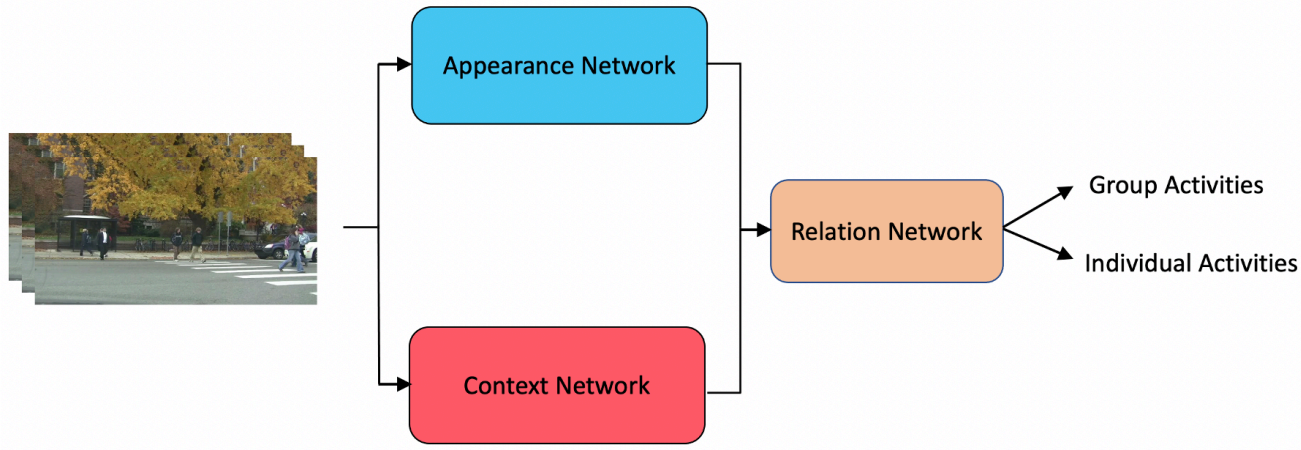
Walking on a **sidewalk**

Crossing a **road**

Key Contributions

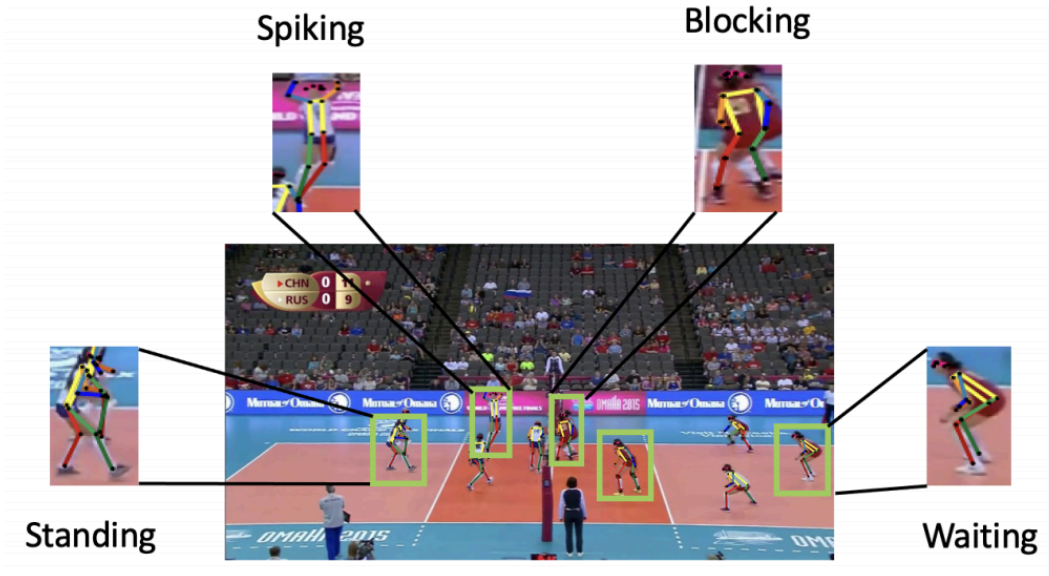
- Leverage contextual cues for group activity
- Two stream network to encode context
- Two types of contextual cues are proposed –
    - Pose
    - Scene Labels

The Proposed Model

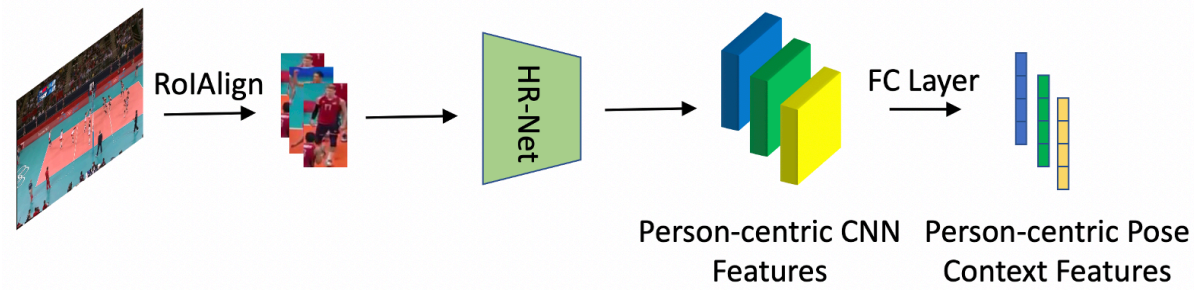


The Pose Contextual Cues

Each activity has its own unique posture

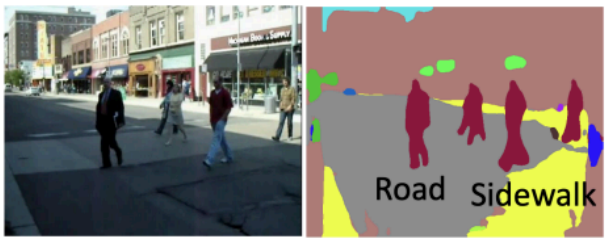


The Pose Context Network



The Scene Contextual Cues

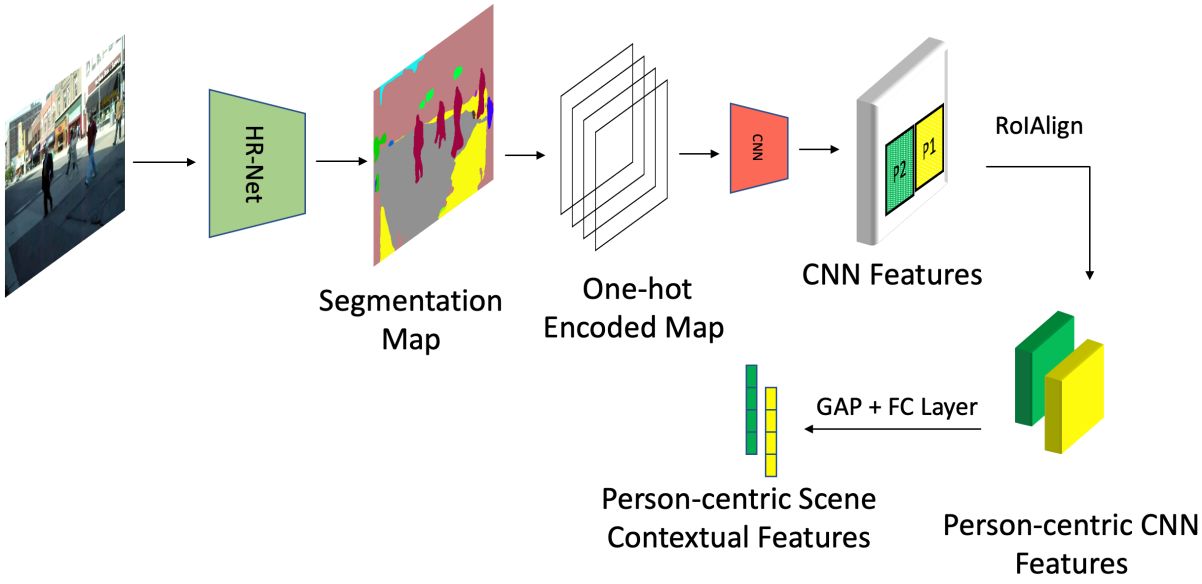
Scene labels provide information about the environment



(a) Crossing activity

(b) Walking activity

The Scene Context Network



Results & Evaluation

Dataset:

- Volleyball –
  - contains 4830 clips of 55 volleyball sports videos
  - 9 individual actions and 8 group activities
- Collective Activity –
  - 44 videos of traffic scenarios
  - 6 individual actions and 5 group activities

Comparison with State-of-the-arts on Volleyball Dataset:

Method	Backbone	Group Activity ↑	Individual Action ↑
Li *et al.*, ICCV'17	Inception-v3	66.90%	-
Ibrahim *et al.*, CVPR'16	AlexNet	81.90%	-
Shu *et al.*, CVPR'17	VGG16	83.30%	-
Biswas *et al.*, WACV'18	AlexNet	83.47%	76.65%
Qi *et al.*, ECCV'18	VGG16	89.30%	-
Ibrahim *et al.*, ECCV'18	VGG19	89.50%	-
Bagautdinov *et al.*, CVPR'17	Inception-v3	90.60%	81.80%
Hu *et al.*, CVPR'20	VGG16	91.4%	-
Wu *et al.*, CVPR'19	Inception-v3	91.62%	81.28%
Azar *et al.*, CVPR'19	I3D	93.04%	-
Ours (Appearance + Pose Context)	Inception-v3 + HR-Net	**93.04%**	**83.02%**

Comparison with State-of-the-arts on Collective Dataset:

Method	Backbone	Group Activity ↑
Lan *et al.*, TPAMI'11	-	79.70%
Choi *et al.*, ECCV'12	-	80.40%
Deng *et al.*, CVPR'16	AlexNet	81.20%
Ibrahim *et al.*, CVPR'16	AlexNet	81.50%
Azar *et al.*, CVPR'19	I3D	85.75%
Li *et al.*, ICCV'17	Inception-v3	86.10%
Shu *et al.*, CVPR'17	VGG16	87.20%
Wu *et al.*, CVPR'19	Inception-v3	88.50%
Wu *et al.*, CVPR'19	VGG19	88.81%
Qi *et al.*, ECCV'18	VGG16	89.10%
Ours (Appearance + Scene Context)	VGG19	**90.07**%

Acknowledgement

This work was supported in part by the ANR AVENUE project (grant ANR-18-CE23-0011). Avijit Dasgupta is supported by a Google India PhD Fellowship.