Action Recognition for Self-Driving Cars

Semester Project (15 ECTS)

Student: Weijiang Xiong (Microengineering)

Supervisor: Prof. Alexandre Alahi, Lorenzo Bertoni, Dr. Taylor Mordan

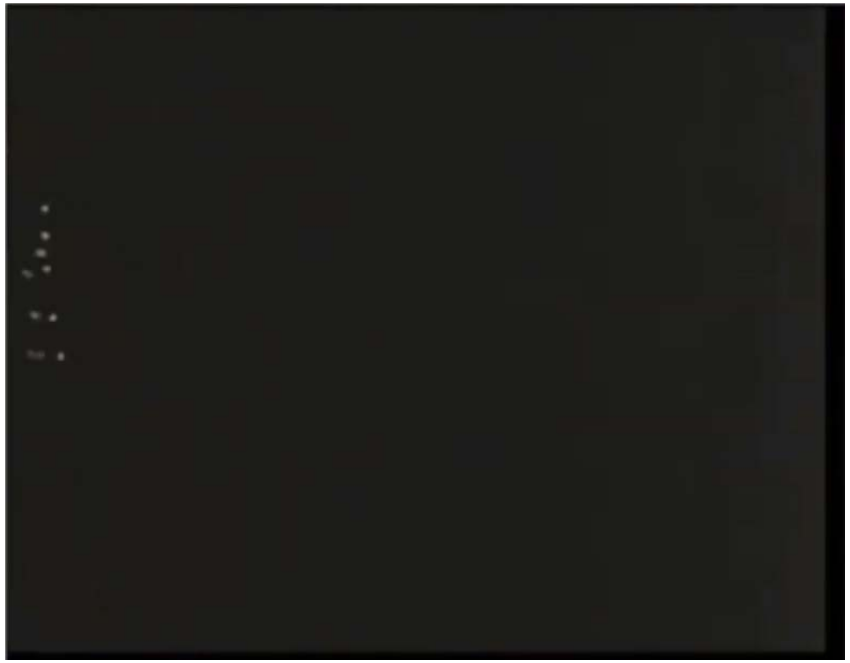
Date: Jan. 12 2022

Image taken from https://usa.honda-ri.com/titan

Introduction and Motivation

Action Recognition

- Identify people's actions in a video/image sequence
- For self-driving cars: understand the environment, make safe decisions and plan reasonable paths



Problem Description

- Input: a video or a sequence of images
- Output: the type of actions for every person inside

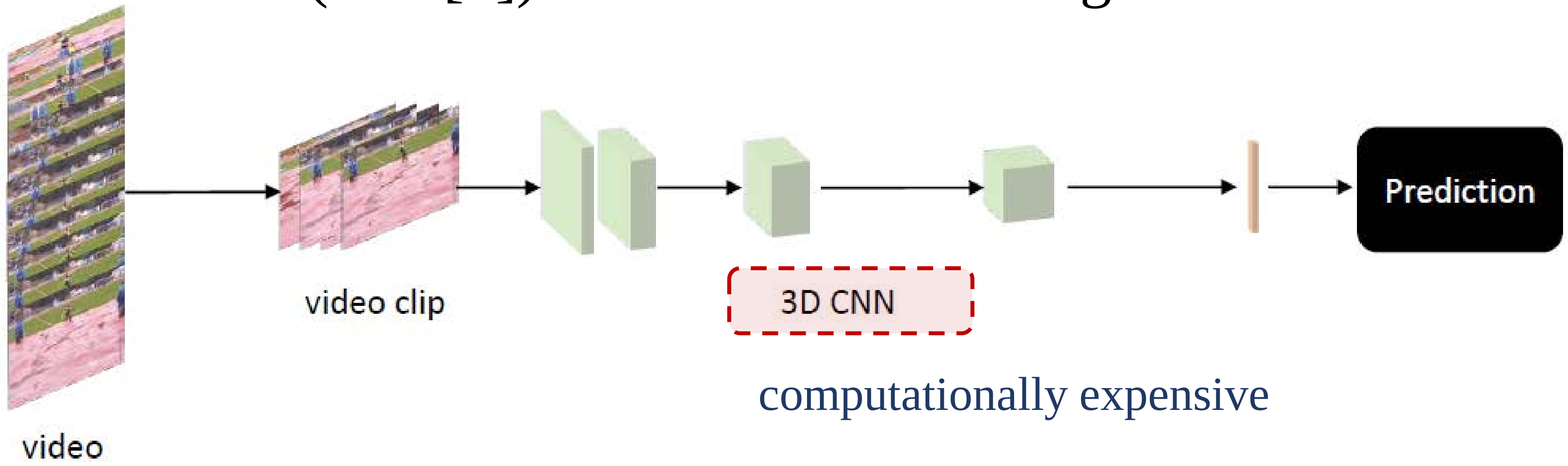
Motivation for Pose-Based Methods

- Human poses are light-weight but highly informative
- Successful work on pose estimation and pose-based vision (related work)

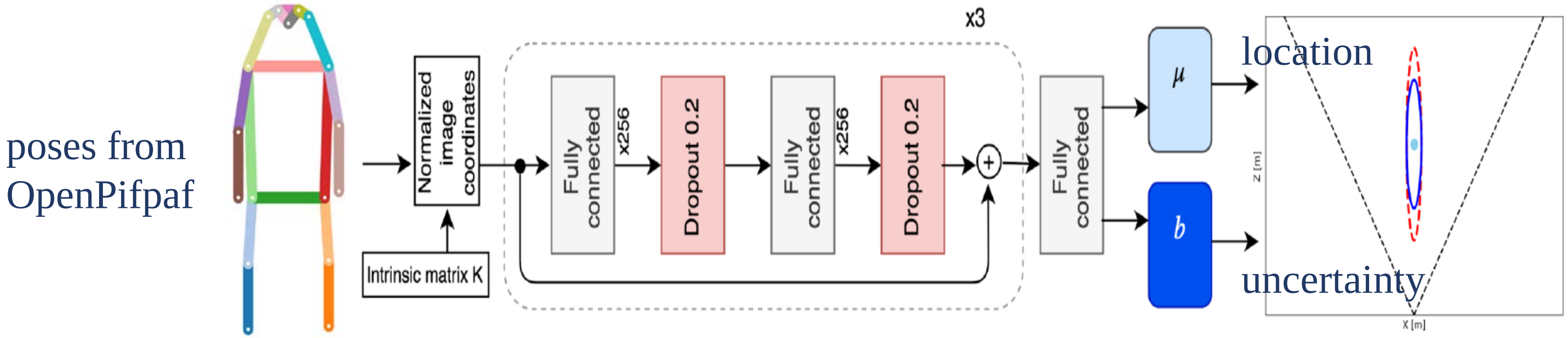
Video Source: Justin Johnson. Umich. Deep Learning for Computer Vision. Lecture 18. 2019 Fall

Related Work

☐ Inflated 3D Convolution (I3D [1]) for video action recognition



☐ OpenPifPaf [2] and MonoLoco [3]



[1] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. CVPR 2017
[2] Sven Kreiss et al., OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. IEEE T-ITS 2021
[3] Lorenzo Bertoni et al., Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. ICCV2019

Datasets and Evaluation

- **TCG [4]: 3D body poses for Traffic Control Gesture**

  550 sequences from different actors (cross-subject evaluation) observed from multiple view points (cross-view)

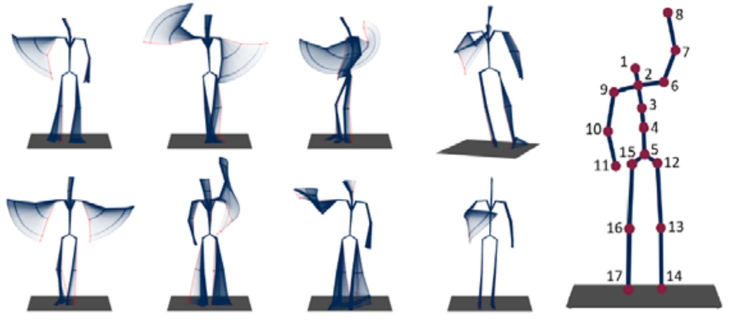
- **TITAN [5]: 700 video clips captured with onboard camera**

  Annotations include five groups of actions, from individual actions (e.g., standing) to those involving context (e.g., talking in group)

- **CASR [6]: Cyclist Arm Signal Recognition**

  178 collected videos, 8 additional videos from youtube (for testing only)

- **Extract poses for TITAN and CASR with OpenPifPaf**



(a) Stop. (b) Clear. (c) Go. (d) Inactive. (e) Skeleton Representation.

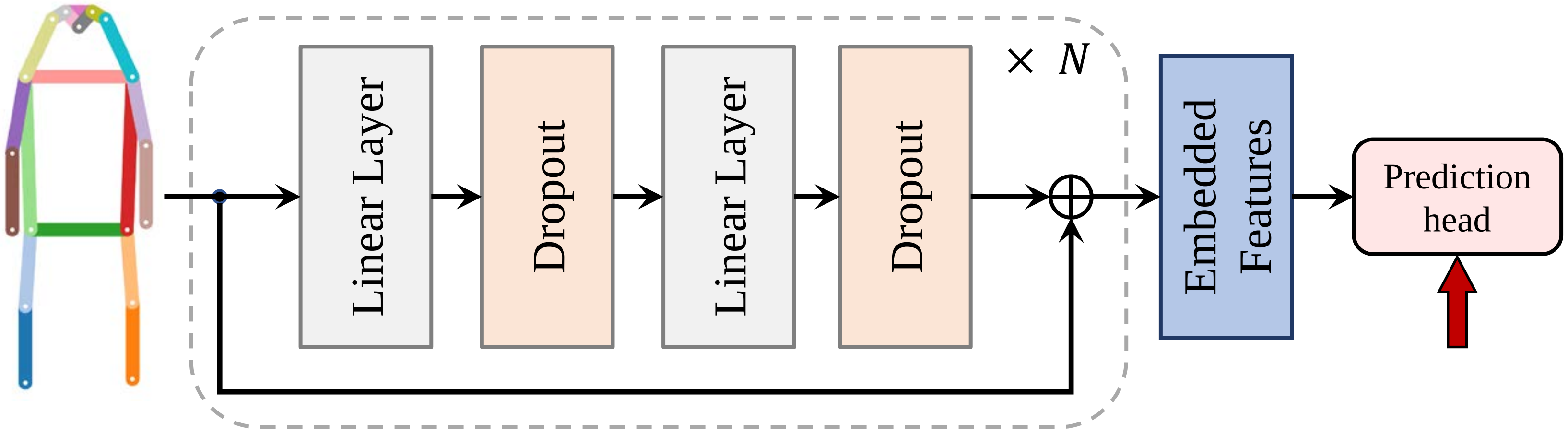
sitting, biking, looking at phone   standing, talking in group

Left, Right, Alternative Right, Stop

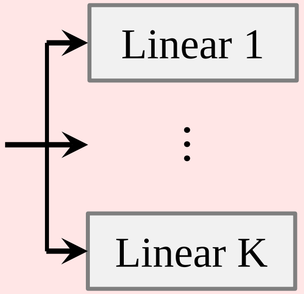
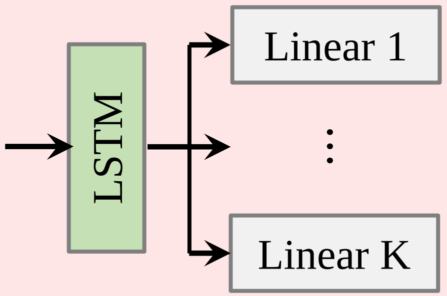
[4] Wiederer et al., Traffic Control Gesture Recognition for Autonomous Vehicles. IROS 2020
[5] Malla Srikanth et al., Titan: Future forecast using action priors. CVPR 2020
[6] Fang et al., Intention Recognition of Pedestrians and Cyclists by 2D Pose Estimation. arXiv:1910.03858

Poseact: Action Recognition with Human Poses

Base Network



Prediction heads

Model Name	Poseact	TempPoseact
Prediction head		
Temporal Info	No	Yes

K=1 for TCG and CASR (usual case)

K=5 (Multitask) for five action groups in TITAN

Action Recognition Results on TCG

Baselines from TCG paper [4]

Method	Cross-subject (%)			Cross-view (%)		
	Accuracy	Jaccard	F1	Accuracy	Jaccard	F1
RNN	82.81	57.40	69.45	80.94	57.21	69.98
GRU	84.44	58.16	70.45	83.47	56.25	68.59
LSTM	83.23	56.32	68.59	79.58	52.02	64.62
Att-LSTM	85.67	50.70	61.87	85.30	59.87	71.20
Bi-GRU	86.80	57.25	68.95	87.37	55.55	67.68
Bi-LSTM	87.24	67.00	78.48	86.66	65.95	77.14
TCN	83.44	62.06	74.23	82.66	63.97	75.95
GCN	65.42	38.55	50.73	62.40	35.05	48.51
AAGCN [7]	91.13	-	85.81	90.22	-	85.21
Pham et al. [7]	91.09	-	86.26	90.64	-	85.52
Poseact	85.03	63.72	76.91	86.29	68.76	80.81
TempPoseact	87.31	69.15	81.15	87.74	70.11	81.89

$$Jaccard = \frac{TP}{TP + FP + FN}$$

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

Temporal info helps, but may not be crucial in TCG

Simple architecture, but still comparable to complicated ones

[4] Wiederer et al., Traffic Control Gesture Recognition for Autonomous Vehicles. IROS 2020
[7] Pham, et al., An Efficient Feature Fusion of Graph Convolutional Networks and Its Application for Real-Time Traffic Control Gestures Recognition. IEEE Access 2021.

Action Recognition Results on TITAN

Classification Accuracy on Test Set

- Poseact uses 2D poses from OpenPifpaf

Results from TITAN paper [5]

Method	I3D	3D ResNet	Poseact (Multitask)
Backbone	InceptionV1	ResNet50	
atomic	0.9219	0.7552	0.8001
simple context	0.5318	0.3173	0.4797
complex context	0.9881	0.9880	0.9780
communicative	0.8649	0.8648	0.8369
transportive	0.9080	0.9081	0.8980
overall	0.8429	0.7667	0.7985

[5] Malla Srikanth et al., Titan: Future forecast using action priors. CVPR 2020

Per-Class Recall (%) of Poseact (Multitask) on TITAN

action group	action type	Rec.	data%
communicative	looking into phone	0	6.05
	talking in group	0	6.99
	talking on phone	0	3.21
	none of the above	**0.999**	**83.76**
complex context	getting in 4 wv	0.018	0.13
	getting off 2 wv	0	0.23
	getting on 2 wv	0	0.12
	getting out of 4 wv	0	0.06
	loading	0	0.20
	unloading	0.046	0.75
	none of the above	**0.992**	**98.50**
atomic	bending	0.362	2.17
	jumping	0	0
	laying down	0	0
	running	0	0.92
	sitting	0.527	4.37
	squatting	0	0.03
	standing	0.129	15.73
	walking	**0.978**	**76.60**

majority class

action group	action type	Rec.	data%
atomic	none of the above	0	0.17
simple context	biking	0.493	3.86
	cleaning an object	0	0.45
	closing	0	0.15
	crossing legally	0.239	7.64
	entering a building	0	0.67
	exiting a building	0.016	0.75
	crossing illegally	0.038	7.22
	motorcycling	0.3	0.09
	opening	0	0.22
	waiting to cross	0.022	1.27
	walking on the side	0.604	35.82
	walking on the road	0.703	25.34
	none of the above	0.269	16.54
transporting	carrying	0.009	6.33
	pulling	0	0.88
	pushing	0.062	2.48
	none of the above	**0.992**	**90.32**

Class Imbalance Problem in TITAN Dataset

☐ Use F1 score as metric

Method	I3D	3D ResNet	Poseact (Multitask)	
Metric	Accuracy	Accuracy	Accuracy	F1
atomic	0.9219	0.7552	0.8001	0.3144
simple	0.5318	0.3173	0.4797	0.1927
complex	0.9881	0.9880	0.9780	0.1529
communicative	0.8649	0.8648	0.8369	0.2278
transportive	0.9080	0.9081	0.8980	0.2634
overall	0.8429	0.7667	0.7985	0.2302

unweighted average over the classes

low score if always predicts the majority class

☐ Focus on a suitable set of actions

- Hard to learn context-dependent actions, especially with insifficient examples (less than 1%)

		walking	standing	sitting	bending	biking	Overall
original annotation	# Instances	43590	10311	304	1297	2057	57559
successful detection	# Detections	32864	6746	189	932	1696	42427
	Percentage	75.4%	65.4%	62.1%	71.8%	82.4%	73.7%

Experiments on Selected Actions

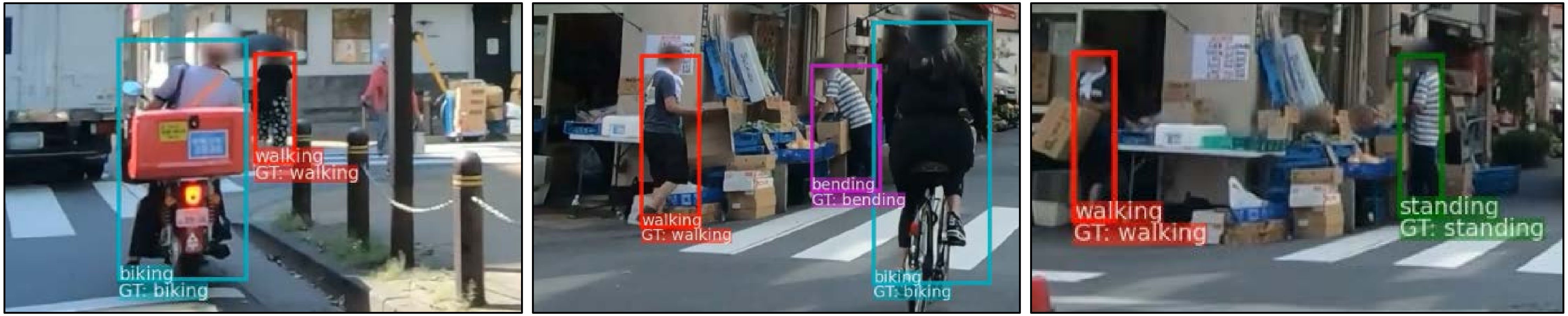
F1 score on the selected actions

Method	walking	standing	sitting	bending	biking	Overall
Poseact (Multitask)	0.884	0.214	0.311	0.457	0.541	0.481
ResNet50 [8]	0.885	0.225	0.01	0.063	0.536	0.344
Poseact	0.919	0.553	**0.771**	**0.621**	**0.839**	**0.741**
TempPoseact	**0.927**	**0.672**	0.710	0.482	0.771	0.712

trained on all 5 action groups
tested on selected actions

trained on selected actions,
not multitask

Recognition examples from Poseact

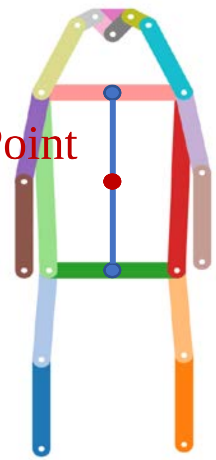


[8] He et al., Deep Residual Learning for Image Recognition. CVPR 2016

Effect of Preprocessing and Temporal Module

Using Relative Keypoint Coordinates

- 17 Absolute coordinates => 1 center location + 17 relative coordinates
- Possible reason: multiple persons in an image, but their actions are not related with their absolute locations, only the body poses matter



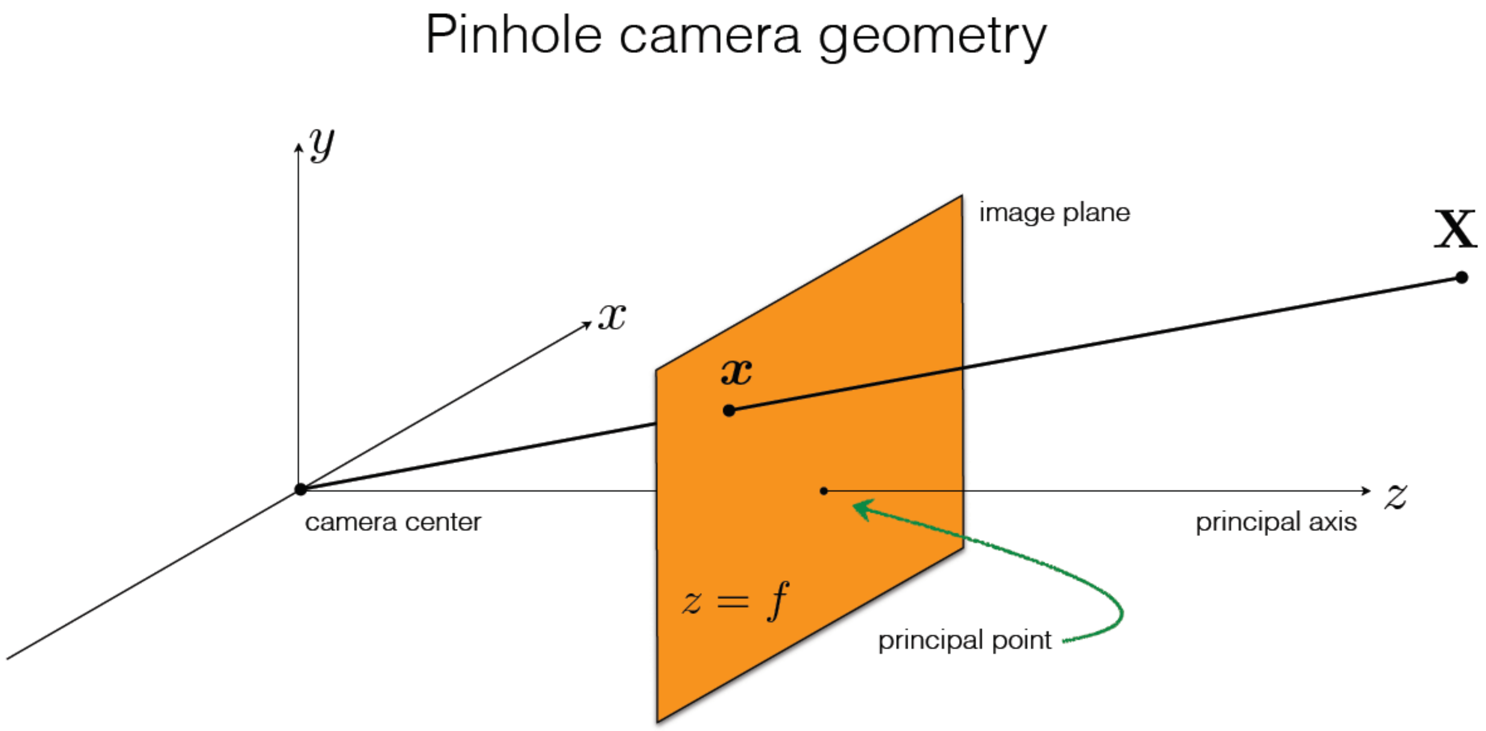
F1 score for each class

- Relative coordinates >> absolute coordinates; center point is not very important
- When using temporal models, we need object track ID to connect new poses to previous ones

Method	Walking	Standing	Sitting	Bending	Biking	**Average**	
Poseact	0.919	0.553	**0.771**	**0.621**	**0.839**	**0.741**	
Abs. Coord	0.887	0.263	0.685	0.478	0.582	0.579	use absolute coordinates
Rm. Center	0.921	0.577	0.742	0.598	0.827	0.733	remove center point
TempPoseact (GT)	**0.927**	**0.672**	0.710	0.482	0.771	0.712	groundtruth object track ID
TempPoseact (PifPaf)	0.923	0.653	0.575	0.467	0.761	0.676	PifPaf object track ID

Explorative Experiments

- Project 3D poses in TCG onto the image plane of a “virtual camera” (acc 20% lower)
  - Possible reason: difficult to choose a proper camera pose to keep the actor in FOV



Picture from CMU 16-385 Computer Vision (Kris Kitani) Lecture 11.1 Camera Matrix

Explorative Experiments

- Project 3D poses in TCG onto the image plane of a “virtual camera” (acc 20% lower)
  - Possible reason: difficult to choose a proper camera pose to keep the actor in FOV
- In TITAN, add phone related actions to selected action set (only ~25% F1)
  - Possible reason: these actions are context-dependent, body poses may not be sufficient

Explorative Experiments

- Project 3D poses in TCG onto the image plane of a “virtual camera” (acc 20% lower)
  - Possible reason: difficult to choose a proper camera pose to keep the actor in FOV
- In TITAN, add phone related actions to selected action set (only ~25% F1)
  - Possible reason: these actions are context-dependent, body poses may not be sufficient
- Apply the models to CASR dataset
  - Good performance on test set, but not on test videos from youtube
  - Possible reason: the dataset is collected seriously (standard arm signal), but people are more casual in real-life (out-of-distribution problem)

Method	Testset		Youtube Set	
	Acc	F1	Acc	F1
Poseact	0.93	0.91	0.68	0.47
TempPoseact	0.92	0.89	0.66	0.46
Heuristics [9]	0.72	0.50	0.79	0.75
Random Forest [6]	0.93	0.92	0.82	0.76

[6] Fang et al., Intention Recognition of Pedestrians and Cyclists by 2D Pose Estimation. arXiv:1910.03858
[9] https://github.com/charlesbvll/monoloco

Discussions

Project Summary

- Learned the recent progress of action recognition
- Created an evaluation code base for TCG, TITAN and CASR
- Experimented basic feed-forward model and temporal model for pose-based action recognition
- Explored preprocessing techniques and generalization applications

Key Takeaways

- A basic feed forward model can recognize non-context-dependent actions from poses
- When multiple persons exist in a frame, it's better to use relative keypoint coordinates
- Take care of the out-of-distribution problem when transferring to different datasets

Future Work

- Advanced models (graph-based, attention mechanism), additional features (speed, angular speed)
- Methods to promote out-of-distribution performance

Github:https://github.com/Weijiang-Xiong/Action_Recognition