Simple Yet Effective Action Recognition for Autonomous Driving

Weijiang Xiong, Lorenzo Bertoni, Taylor Mordan* and Alexandre Alahi

École Polytechnique Fédérale de Lausanne (EPFL)
Visual Intelligence for Transportation (VITA)
CH-1015 Lausanne, Switzerland
firstname.lastname@epfl.ch, * Corresponding author

Extended abstract submitted for presentation at the 11th Triennial Symposium on Transportation Analysis conference (TRISTAN XI)
June 19-25, 2022, Mauritius Island

March 31, 2022

Keywords: Action Recognition, Autonomous Driving, Deep Learning, Human Pose Estimation

1 INTRODUCTION

Self-driving cars and delivery robots are set to shape the future of transportation, but they still have to learn how to co-exist with humans in close proximity. Autonomous systems need to detect pedestrians and understand the meaning of their actions before making appropriate decisions in response. Action recognition is therefore an essential task for transportation applications, and yet very challenging, as there is no control over the distances of pedestrians or the real-world variations like lighting, weather, and occlusions.

In this paper, we focus on the action recognition task in the context of transportation applications and deal with real-world variations and challenging scenarios by representing humans through their 2D poses. Human poses are an effective intermediate representation for 2D and 3D human perception tasks. Representing human postures as sparse sets of keypoints allows focusing on essential details while providing invariance to many factors, including background scenes, lighting, textures, and clothes. Methods leveraging keypoints have obtained state-of-the-art results and excellent generalization properties on 3D pedestrian localization (Bertoni *et al.*, 2019) and 3D pose estimation (Martinez *et al.*, 2017) tasks. However, keypoints' greatest strength is also their main weakness, as such a low-dimensional representation risks neglecting other essential elements in a scene. We propose a simple approach using keypoints as intermediate representations and aim to shed light on which tasks keypoints are effective representations for. We first validate our approach on the TCG dataset (Wiederer *et al.*, 2020), showing that a simple method can achieve better results than temporal baselines using LSTMs, and comparable results with complex attention-based graph convolutional networks. Then, we compare our approach on the action recognition task on TITAN (Malla *et al.*, 2020), a new dataset for autonomous driving. We show that on atomic actions, such as *walking* or *standing*, our keypoint-based approach outperforms an image-based method, validating the effectiveness of human poses as intermediate representations for action recognition tasks in transportation applications. We publicly share our source code to facilitate academic communication.[1]

2 PROPOSED APPROACH

We propose a simple method (Figure 1) to recognize human activities from images, leveraging human poses as intermediate representations. This model is inspired by MonoLoco (Bertoni

[1] Github repo: https://github.com/vita-epfl/pose-action-recognition

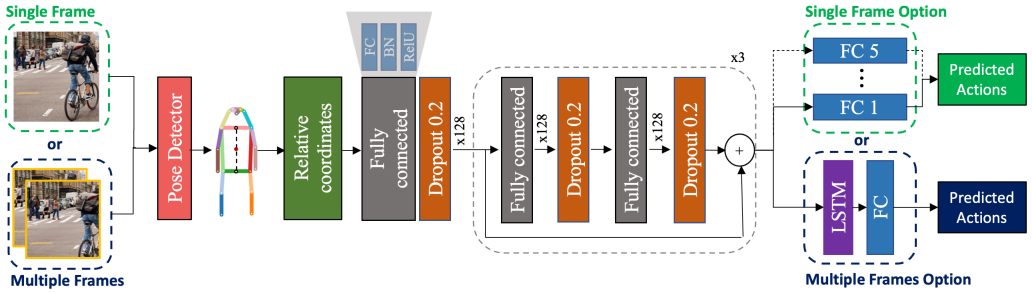

Figure 1 – *Model architecture inspired by MonoLoco. The input is a set of 2D joints extracted from a raw image and the output is the estimated action of a pedestrian. We use three different heads for image-based and video-based action recognition, and for estimating simultaneous actions.*

Table 1 – *Action recognition results on TCG (Wiederer et al., 2020) test set.*

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	91.13	-	85.81	90.22	-	85.21
Pham *et al.*	91.09	-	86.26	90.64	-	85.52
Ours (single-frame)	85.03	63.72	76.91	86.29	68.76	80.81
Ours (temporal)	87.31	69.15	81.15	87.74	70.11	81.89

et al., 2019) and Figure 1 presents the general network architecture, consisting of two stages. We first extract 17 body keypoints for each person in an image with OpenPifPaf (Kreiss *et al.*, 2021), and transform them into 1 center point and 17 relative coordinates (18 in total, shown in Figure 1). Then, the model encodes the keypoints with a feedforward network, and predicts the corresponding action from the encoded representation. We also extend our method (i) to estimate actions from videos and (ii) to predict simultaneous groups of actions (*e.g.*, a person can walk while being on the phone). For (i), we add a simple LSTM to process a temporal sequence of poses before the final linear layer, and for (ii), we use a multitask approach, where multiple parallel heads (instead of just one) process a shared representation to yield multiple predictions.

3 EXPERIMENTS

3.1 Experiments on TCG

The TCG dataset (Wiederer *et al.*, 2020) collects accurate 3D body keypoints for recognition of traffic control gestures, which we use to validate the design of our single-frame model directly from error-free poses. Following the cross-subject and cross-view evaluation protocols in TCG, Table 1 compares the performances of our single-frame and temporal models with eight simple baseline methods as well as two more complex attention-based graph convolutional networks (Pham *et al.*, 2021). Our temporal model outperforms the simple baseline models. Specifically, it performs better than the LSTM baseline, which directly predicts actions from raw keypoint coordinates.

Table 2 – *Action recognition results on TITAN (Malla et al., 2020) test set.*

	I3D	3D ResNet	Ours (multitask)	
Action group	Accuracy (%)	Accuracy (%)	Accuracy (%)	mAP (%)
atomic	92.19	75.52	80.01	26.80
simple	53.18	31.73	47.97	20.27
complex	98.81	98.80	97.80	15.50
communicative	86.49	86.48	83.69	29.55
transportive	90.80	90.81	89.80	28.30
overall	84.29	76.67	79.85	24.08

Table 3 – *Action recognition results with selected actions on TITAN (Malla et al., 2020) test set.*

		Average Precision (AP %) ↑ [Detection Recall ↑]					
Method	Inputs	Walking [75.4%]	Standing [65.4%]	Sitting [62.1%]	Bending [71.8%]	Biking [82.4%]	Average [73.7%]
Ours (multitask)	Keypoints	90.16	40.67	43.78	41.12	57.70	48.14
ResNet50 (He *et al.*)	Crops	92.85	42.07	5.18	8.44	56.00	40.91
Ours (single-frame)	Keypoints	96.87	64.55	81.22	64.59	88.30	79.11
Ours (temporal)	Keypoints	97.83	73.02	65.78	47.31	84.98	73.78

This demonstrates the effectiveness of processing raw keypoints with a feedforward network. Our models are still outperformed by the two attention-based graph convolutional networks, but those are much heavier and would likely be less suitable for applications with hard run-time constraints (e.g., memory footprint, inference time), such as autonomous driving. Additionally, traffic control gestures are designed to be unambiguous actions that could easily be understood without temporal context, which means temporal information should not be crucial for these gestures. The close results of our single-frame and temporal models confirm this observation.

3.2 Experiments on TITAN

The TITAN dataset (Malla *et al.*, 2020) has 700 video clips captured by an on-board camera, which are suitable for evaluating the complete recognition workflow starting from raw images. All annotated actions belong to *atomic*, *simple context*, *complex context*, *communicative* or *transportive* action groups. Notably, all the people in all the frames are annotated with five action labels, i.e., one from each action group (including labels for *no action* for any group).

In Table 2, we compare our multitask model, using five prediction heads to match the five action groups of TITAN, with I3D (Carreira & Zisserman, 2017) and 3D ResNet (Hara *et al.*, 2018). We observe it has comparable accuracy to the other two methods. However, TITAN is highly imbalanced (toward *no action* for almost all groups), thus the overall accuracy is not a suitable metric to evaluate results. For this reason, we introduce mean Average Precision (mAP), where Average Precisions (APs) are computed for all classes separately and then averaged.

However, since the labels *no action* dominate most action groups, and some actions have insufficient numbers of examples, we focus on a subset of actions where our multitask model has reasonable mAP. We also merge actions *biking* and *motorcycling* with close meaning, resulting in five classes: *walking*, *standing*, *sitting*, *bending* and *biking*. Table 3 compares the recognition performances of four models using mAP. Our multitask model is trained on the original TITAN dataset, and we only keep the predictions corresponding to the selected action subset. Since the original dataset contains considerable *no action* samples, the training process of this model is dominated by this majority class, and the model does not have satisfactory mAP. The follow-