Keypoint-based Activity Recognition for Autonomous Driving

Weijiang Xiong[a], Lorenzo Bertoni[a,*], Taylor Mordan[a] and Alexandre Alahi[a]

[a] École Polytechnique Fédérale de Lausanne (EPFL)
Visual Intelligence for Transportation (VITA)
CH-1015 Lausanne, Switzerland
weijiang.xiong@epfl.ch, lorenzo.bertoni@epfl.ch, taylor.mordan@epfl.ch, alexandre.alahi@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

January 14, 2022

Keywords: Action Recognition, Autonomous Driving, Semantic Keypoints, Visual Perception

1 INTRODUCTION

Semantic keypoints (SM) are a popular representation for 2D and 3D human perception tasks. They allow to focus on essential details on human postures while providing invariance to many factors, including background scenes, lighting, textures and clothes. SMs can be used as a low-dimensional, intermediate representation for alternative pipeline to end-to-end networks leveraging raw images. A recent state-of-the-art keypoint estimator (Kreiss *et al.*, 2021) has enabled various high-quality methods in various vision tasks, such as 3D localization (Bertoni *et al.*, 2019). However, SMs greatest strength is also their main weakness, as such a low-dimensional representation is at risk of neglecting other essential elements in a scene.
Human action recognition has been an active research topic with the recent advancements in deep-learning-based video understanding. Carreira & Zisserman propse to inflate the architecture of a pretrained CNN, and generalize the input from images to videos. But such a model requires considerable computation resources. Pham *et al.* proposed an efficient feature fusion mechanism to enhance the AAGCN (Shi *et al.*, 2020), and applied the method to action recognition task.
In this work, we focus on the action recognition task and review the effectiveness of keypoint-based method on different datasets. We propose a simple model and compare it with other keypoints-based ones on TCG (Wiederer *et al.*, 2020). We use TITAN (Malla *et al.*, 2020) to compare on various action groups, and we find our model is suitable for several basic actions, after applying proper preprocessing and addressing the class imbalance problem.

2 PROPOSED APPROACH

We propose a concise and effective keypoint-based method, i.e., Poseact, to recognize human activities from images, with body poses as intermediate representations. This model is inspired by MonoLoco (Bertoni *et al.*, 2019) and Figure 1 presents the general network architecture of Poseact. In general, our method consists of two stages. We first extract the human body keypoints from images with OpenPifPaf (Kreiss *et al.*, 2021), and we transform the 17 keypoints into 1 center point and 17 relative coordinates (18 in total). After that, the model takes the keypoints, encode them with a series of feedforward blocks, and predict the action with the embedded features. Our experiments involves two variants of Poseact. The temporal variant, i.e., TempPoseact, has an LSTM layer to process the embedded features immediately before the final linear layer. Meanwhile, the multi-head variant, i.e., Poseact (Multihead), has multiple linear

layers to produce multiple sets of classification results based on the shared embedded features. The implementation of this work has been published to facilitate academic communication[1].

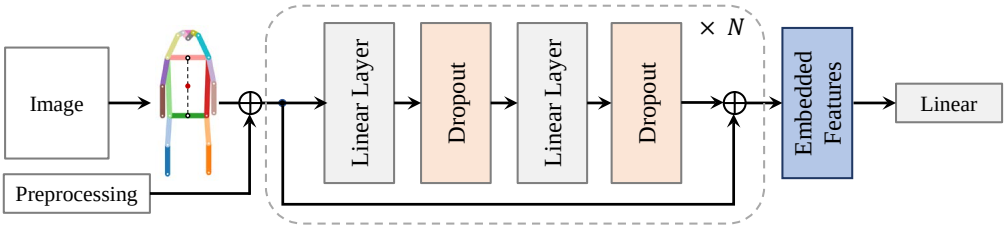

Figure 1 – *General structure of Poseact. Center point in red, points with black edges are auxiliary.*

3 EXPERIMENTS

3.1 Experiments on TCG

The TCG dataset (Wiederer *et al.*, 2020) provides 3D body keypoints for traffic control gesture, and the dataset utilizes cross-subject and cross-view evaluation protocol to cover the variation in actors and view points. Table 1 compares the performance of Poseact with eight baseline methods in TCG paper as well as two AAGCN-based methods introduced in (Pham *et al.*, 2021).
Our temporal model, TempPoseact, shows better results than all the eight baseline models. Specifically, it's much better than the LSTM baseline, which directly predicts actions from raw keypoint coordinates. On the contrary, TempPoseact first encodes the keypoint coordinates into intermediate features, and then classify the action based on these features. Therefore, we observe that the embedded feature are more suitable for classification task than the raw keypoints.
Meanwhile, the classification accuracy of the single frame model, Poseact, is quite close to TempPoseact, which means on TCG, temporal information can help improve the performance, but it's not crucial for the action recognition task. We deduce that these actions are designed to be unambiguous, therefore one could understand their meanings without continuous observation.
Despite the fact that Poseact has a straightforward architecture, its performance is still comparable to the complicated AAGCN-based models. Therefore we believe Poseact should be more suitable for applications that has limited computational resources, such as the on-board perception system of a self-driving cars.

Table 1 – *Action Recognition Results on TCG*

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
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

[1] Github repo: https://github.com/Weijiang-Xiong/Action_Recognition

3.2 Experiments on TITAN

The TITAN dataset (Malla *et al.*, 2020) has 700 video clips captured by an on-board camera, and the annotated actions consists of five groups, i.e., atomic, simple context, complex context, communicative and transportive actions. Notably, in each frame, each person is annotated with all these five action groups. For example, a person could be both sitting (atomic) and looking into phone (communicative). if the person's action is not considered by the dataset, his/her action will be annotated as "none of the above". Therefore, we first developed a multi-head variant of Poseact to match the five action groups.
Malla *et al.* have evaluated I3D (Carreira & Zisserman, 2017) and 3D ResNet (Hara *et al.*, 2018) on TITAN, and Table 2 compares the multi-head Poseact with these two methods. We observe that Poseact has similar accuracy, compared to the other two methods, and the overall accuracy is even higher than 3D ResNet. However, we found out the TITAN dataset is highly imbalanced, and thus the overall accuracy is not a suitable metric to evaluate these methods. For example, in the complex context action group, more than 98% of the persons are labeled with "none of the above" and therefore a model can always predict this majority class and have nearly perfect accuracy. As such, we introduce the mean average precision (mAP) as the main performance metric. mAP is an unweighted average over the classes, and therefore the model can not obtain high mAP by always predicting the majority class.

Table 2 – *Recognition performance on the original TITAN dataset*

Method	I3D	3D ResNet	Poseact (Multitask)	
Metric	Accuracy	Accuracy	Accuracy	mAP
atomic	0.9219	0.7552	0.8001	0.2680
simple	0.5318	0.3173	0.4797	0.2027
complex	0.9881	0.9880	0.9780	0.1550
communicative	0.8649	0.8648	0.8369	0.2955
transportive	0.9080	0.9081	0.8980	0.2830
overall	0.8429	0.7667	0.7985	0.2408

With mAP, we can reasonably evaluate the performance on the imbalanced dataset. However, the majority class could still easily overwhelm the training process, and thus we have to focus on a suitable set of actions. Considering the fact that "none of the above" is not informative, and some actions in TITAN dataset have insufficient number of examples, we filtered the original annotations and focused on a suitable set of actions. Precisely, we select five actions from the original annotations, i.e., walking, standing, sitting, bending and biking. Notably, biking belongs to the simple context action group, and the others are considered to be atomic. Therefore, when filtering the dataset, we give priority to biking, and then consider the atomic actions, and in our experiments, biking includes motorcycling as well. For example, if a person is annotated as sitting and motorcycling, motorcycling will be considered as biking, and then precede sitting. As a result, the person is considered to be biking in the selected action set.
Table 3 compares the recognition performance of four models using mAP. Poseact (Multitask) is trained on the original TITAN dataset, and we transform its prediction on test set into the selected set with the same procedure specified above. Since the original dataset contains considerable "none of the above" samples, the training process of this model could have been dominated by the majority class, and thus the model does not have satisfactory mAP.
The following three models are trained and tested on the selected action set. We crop an image patch around a detected person, and train a ResNet50 (He *et al.*, 2016) on these image crops. Following Malla *et al.*, we associate the detected poses with ground truth object track ID. Then, we train and evaluate TempPoseact with the obtained sequences. The results show that TempPoseact is better at walking and standing, while the single frame Poseact is better for

sitting, bending and biking. Figure 2 presents several qualitative examples.

Table 3 – *Recognition performance on selected actions*

Method	Input	Average Precision (AP) ↑					
		Walking	Standing	Sitting	Bending	Biking	Average
Poseact (Multitask)	Keypoints	0.9016	0.4067	0.4378	0.4112	0.5770	0.4814
ResNet50 (He *et al.*)	Crops	0.9285	0.4207	0.0518	0.0844	0.5600	0.4091
Poseact	Keypoints	0.9687	0.6455	**0.8122**	**0.6459**	**0.8830**	**0.7911**
TempPoseact	Keypoints	**0.9783**	**0.7302**	0.6578	0.4731	0.8498	0.7378
Detection Recall ↑	-	75.4%	65.4%	62.1%	71.8%	82.4%	73.7%

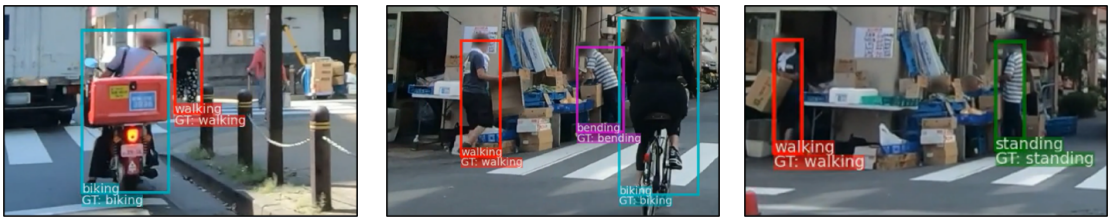

Figure 2 – *Recognition Examples*

In ablation study, we further evaluate the effect of preprocessing and LSTM module. If we disable the preprocessing, and use absolute keypoint coordinates, the mAP degrades from 0.741 to 0.579, but the performance remains stable without the center point. A possible reason is, when multiple persons exists in a frame, their actions are reflected by the movements of their body, and the absolute location is irrelevant. For the TempPoseact, we utilized the annotated object track ID to associate the poses across time, and the mAP degrades from 0.712 to 0.676 if we switch to the track ID from OpenPifPaf.

Table 4 – *Average Precision on the selected action set*

Method	Walking	Standing	Sitting	Bending	Biking	mAP
Poseact	0.919	0.553	**0.771**	**0.621**	**0.839**	**0.741**
Absolute Coordinate	0.887	0.263	0.685	0.478	0.582	0.579
Remove Center Point	0.921	0.577	0.742	0.598	0.827	0.733
TempPoseact (GT)	**0.927**	**0.672**	0.710	0.482	0.771	0.712
TempPoseact (PifPaf)	0.923	0.653	0.575	0.467	0.761	0.676

4 Conclusion

In this work, we present Poseact, a concise and effective method to recognize human actions with 2D poses estimated from images. We show that with proper preprocessing, the single frame Poseact could classify basic human actions with satisfactory precision, and its performance could be comparable to some complicated models. We believe Poseact would facilitate the application of real-time perception for autonomous vehicles.

References

Bertoni, Lorenzo, Kreiss, Sven, & Alahi, Alexandre. 2019 (October). MonoLoco: Monocular 3D Pedestrian Localization and Uncertainty Estimation. *In: the IEEE International Conference on Computer Vision (ICCV).*

Carreira, Joao, & Zisserman, Andrew. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. *Pages 6299–6308 of: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hara, Kensho, Kataoka, Hirokatsu, & Satoh, Yutaka. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? *Pages 6546–6555 of: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, & Sun, Jian. 2016. Deep residual learning for image recognition. *Pages 770–778 of: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kreiss, Sven, Bertoni, Lorenzo, & Alahi, Alexandre. 2021. OpenPifPaf: Composite Fields for Semantic Keypoint Detection and Spatio-Temporal Association. *IEEE Transactions on Intelligent Transportation Systems*, March, 1–14.

Malla, Srikanth, Dariush, Behzad, & Choi, Chiho. 2020. Titan: Future forecast using action priors. *Pages 11186–11196 of: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Pham, Dinh-Tan, Pham, Quang-Tien, Le, Thi-Lan, & Vu, Hai. 2021. An Efficient Feature Fusion of Graph Convolutional Networks and Its Application for Real-Time Traffic Control Gestures Recognition. *IEEE Access*, **9**, 121930–121943.

Shi, Lei, Zhang, Yifan, Cheng, Jian, & Lu, Hanqing. 2020. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing*, **29**, 9532–9545.

Wiederer, Julian, Bouazizi, Arij, Kressel, Ulrich, & Belagiannis, Vasileios. 2020. Traffic Control Gesture Recognition for Autonomous Vehicles. *Pages 10676–10683 of: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE.