Research Internship:
3D Pose Estimation for Driver Monitoring

Victoria BRAMI

Master Mathématiques Vision Apprentissage (MVA)

victoria.brami@eleves.enpc.fr

Supervised by Patrick Pérez, advised by Souhaiel Khalfaoui and Renaud Marlet

Thursday September 29^{th} 2022

École des Ponts ParisTech

valeo.ai

Context of work

- Distraction accounts for **20%** of car accidents in 2020.[1]
- **Driver Monitoring System (DMS)**: Set of equipment tools developed around the driver to ease his way of driving.
- EU Comission: new regulations on DMS to be introduced by 2024.

→ Necessity to Improve existing systems.

École des Ponts ParisTech

valeo.ai

[1] Report made by the European Comission in 2020.

Context of work

Motivations:

Get knowledge of in-car occupation to understand the occupants' behaviour while driving.

Supply the best IMS possible (security, confort, etc.)

Our Goal:

Propose **a real-time 3D Pose Estimation of the driver** to be capable to analyse his activities in a second phase.

Outline

Interior Monitoring Datasets Constraints

Dataset	D&A [5]	Pandora[1]	AutoPOSE[7]	TICaM[3]	DMD[6]	DAD[4]	DriPE[2]
Scene Type	Real Condition	Sitting, Driving-like	In-Cabin Driving	In-Cabin Driving	Real Condition	Real Condition	Real Condition
Occupants	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only
Views	6	1	2	1	3	2	>1
Nb. frames	>9.6M	250k	1.1M / 315k(view 1)	119.7k / 3.3k	4.4M	2.1M	10k
Nb. videos	29	110	21			386	-
RGB/Gray	✓	✓	✓	✓	✓	-	✓
IR	✓	-	✓	✓(6.7k)	✓	✓	-
Depth	✓	✓	✓	✓(6.7k)	✓	✓	-
Subjects[a]	15 (4/11)	22 (10/12)	21 (10/11)	13 (N/A)	37 (10/27)	31 (N/A)	19 (7/12)
Annotations Contents							
Dataset	D&A [5]	Pandora[1]	AutoPOSE[7]	TICaM[3]	DMD[6]	DAD[4]	DriPE[2]
Activity	✓	✓	-	✓	✓	-	-
Nb. Activ.	83	20	-	20	13	-	-
2D joints	✓	✓	-	✓	-	-	✓
3D joints	✓	✓	✓	-	N/A	-	-
Format	COCO 17	17 Upper	Head center	COCO 17	-	-	COCO17

Table: Main large-scale Driver Monitoring datasets

[a](F/M) for female / male

Interior Monitoring Datasets Constraints

Dataset	D&A [5]	Pandora[1]	AutoPOSE[7]	TICaM[3]	DMD[6]	DAD[4]	DriPE[2]
Scene Type	Real Condition	Sitting, Driving-like	In-Cabin Driving	In-Cabin Driving	Real Condition	Real Condition	Real Condition
Occupants	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only
Views	6	1	2	1	3	2	>1
Nb. frames	>9.6M	250k	1.1M / 315k(view 1)	119.7k / 3.3k	4.4M	2.1M	10k
Nb. videos	29	110	21			386	-
RGB/Gray	✓	✓	✓	✓	✓	-	✓
IR	✓	-	✓	✓(6.7k)	✓	✓	-
Depth	✓	✓	✓	✓(6.7k)	✓	✓	-
Subjects[a]	15 (4/11)	22 (10/12)	21 (10/11)	13 (N/A)	37 (10/27)	31 (N/A)	19 (7/12)
Annotations Contents							
Dataset	D&A [5]	Pandora[1]	AutoPOSE[7]	TICaM[3]	DMD[6]	DAD[4]	DriPE[2]
Activity	✓	✓	-	✓	✓	-	-
Nb. Activ.	83	20	-	20	13	-	-
2D joints	✓	✓	-	✓	-	-	✓
3D joints	✓	✓	✓	-	N/A	-	-
Format	COCO 17	17 Upper	Head center	COCO 17	-	-	COCO17

Table: Main large-scale Driver Monitoring datasets

École des Ponts ParisTech

valeo.ai

[a] (F/M) for female / male

Interior Monitoring Datasets Constraints

Dataset	D&A [5]	Pandora[1]	AutoPOSE[7]	TICaM[3]	DMD[6]	DAD[4]	DriPE[2]
Scene Type	Real Condition	Sitting, Driving-like	In-Cabin Driving	In-Cabin Driving	Real Condition	Real Condition	Real Condition
Occupants	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only
Views	6	1	2	1	3	2	>1
Nb. frames	>9.6M	250k	1.1M / 315k(view 1)	119.7k / 3.3k	4.4M	2.1M	10k
Nb. videos	29	110	21			386	-
RGB/Gray	✓	✓	✓	✓	✓	-	✓
IR	✓	-	✓	✓(6.7k)	✓	✓	-
Depth	✓	✓	✓	✓(6.7k)	✓	✓	-
Subjects[a]	15 (4/11)	22 (10/12)	21 (10/11)	13 (N/A)	37 (10/27)	31 (N/A)	19 (7/12)
Annotations Contents							
Dataset	D&A [5]	Pandora[1]	AutoPOSE[7]	TICaM[3]	DMD[6]	DAD[4]	DriPE[2]
Activity	✓	✓	-	✓	✓	-	-
Nb. Activ.	83	20	-	20	13	-	-
2D joints	✓	✓	-	✓	-	-	✓
3D joints	✓	✓	✓	-	N/A	-	-
Format	COCO 17	17 Upper	Head center	COCO 17	-	-	COCO17

Table: Main large-scale Driver Monitoring datasets

École des Ponts ParisTech

valeo.ai

[a] (F/M) for female / male

Interior Monitoring Datasets Constraints

Dataset	D&A [5]	Pandora[1]	AutoPOSE[7]	TICaM[3]	DMD[6]	DAD[4]	DriPE[2]
Scene Type	Real Condition	Sitting, Driving-like	In-Cabin Driving	In-Cabin Driving	Real Condition	Real Condition	Real Condition
Occupants	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only
Views	6	1	2	1	3	2	>1
Nb. frames	>9.6M	250k	1.1M / 315k(view 1)	119.7k / 3.3k	4.4M	2.1M	10k
Nb. videos	29	110	21			386	-
RGB/Gray	✓	✓	✓	✓	✓	-	✓
IR	✓	-	✓	✓(6.7k)	✓	✓	-
Depth	✓	✓	✓	✓(6.7k)	✓	✓	-
Subjects[a]	15 (4/11)	22 (10/12)	21 (10/11)	13 (N/A)	37 (10/27)	31 (N/A)	19 (7/12)
Annotations Contents							
Dataset	D&A [5]	Pandora[1]	AutoPOSE[7]	TICaM[3]	DMD[6]	DAD[4]	DriPE[2]
Activity	✓	✓	-	✓	✓	-	-
Nb. Activ.	83	20	-	20	13	-	-
2D joints	✓	✓	-	✓	-	-	✓
3D joints	✓	✓	✓	-	N/A	-	-
Format	COCO 17	17 Upper	Head center	COCO 17	-	-	COCO17

Table: Main large-scale Driver Monitoring datasets

[a](F/M) for female / male

Interior Monitoring Datasets Constraints

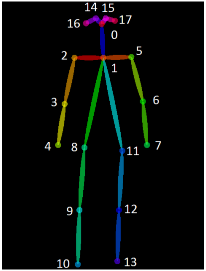
Dataset	D&A [5]	Pandora[1]	AutoPOSE[7]	TICaM[3]	DMD[6]	DAD[4]	DriPE[2]
Scene Type	Real Condition	Sitting, Driving-like	In-Cabin Driving	In-Cabin Driving	Real Condition	Real Condition	Real Condition
Occupants	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only	Driver Only
Views	6	1	2	1	3	2	>1
Nb. frames	>9.6M	250k	1.1M / 315k(view 1)	119.7k / 3.3k	4.4M	2.1M	10k
Nb. videos	29	110	21			386	-
RGB/Gray	✓	✓	✓	✓	✓	-	✓
IR	✓	-	✓	✓(6.7k)	✓	✓	-
Depth	✓	✓	✓	✓(6.7k)	✓	✓	-
Subjects[a]	15 (4/11)	22 (10/12)	21 (10/11)	13 (N/A)	37 (10/27)	31 (N/A)	19 (7/12)
Annotations Contents							
Dataset	D&A [5]	Pandora[1]	AutoPOSE[7]	TICaM[3]	DMD[6]	DAD[4]	DriPE[2]
Activity	✓	✓	-	✓	✓	-	-
Nb. Activ.	83	20	-	20	13	-	-
2D joints	✓	✓	-	✓	-	-	✓
3D joints	✓	✓	✓	-	N/A	-	-
Format	COCO 17	17 Upper	Head center	COCO 17	-	-	COCO17

Table: Main large-scale Driver Monitoring datasets

[a](F/M) for female / male

Drive And Act Dataset Format

- 6 views.
- 15 drivers filmed 20-30 min each (10 / 2 / 3).
- **9.6 Million** frames.
- Annotations triangulated from **OpenPose**[2]



(a) COCO17 annotation format[4]

(b) Sample from Drive&Act[5]

[2] Cao & al., OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, in *TPAMI*, 2019.

[4] Lin & al., Microsoft COCO: Common objects in context, in *ECCV*, 2014.

[5] Martin & al., Drive&Act: A Multi-modal Dataset for Fine-grained Driver Behavior Recognition in Autonomous Vehicles, in *ICCV*, 2019.

Outline

2D Pose Models Fields

In HR Net (Top Down)

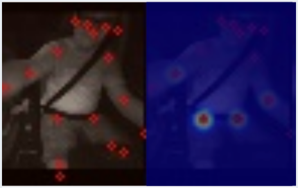
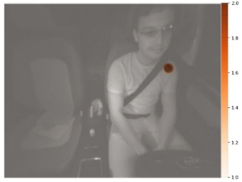


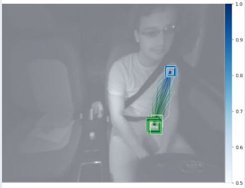
Figure: Heatmap

In OpenPifPaf (Bottom-Up)

(a) CIF



(b) CAF



École des Ponts ParisTech

valeo.ai

Top Down Model: HR-Net (2019)

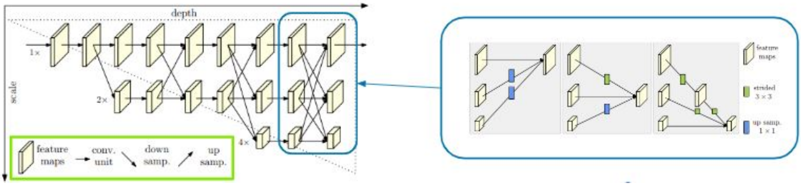


Figure: HR-Net Model Architecture[6]

[6] Sun & al., Deep High-Resolution Representation Learning for Human Pose Estimation, in *CVPR*, 2019.

Bottom-Up Model: OpenPifPaf (2019-2021)

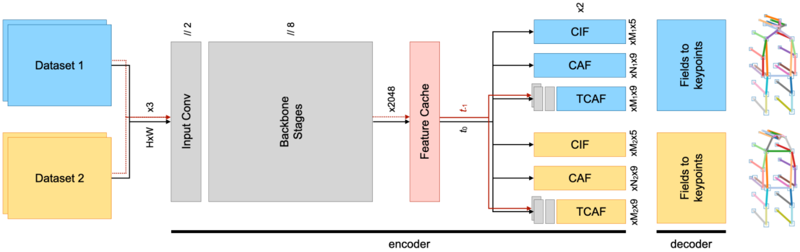


Figure: OpenPifPaf Model Architecture[7]

[7] Kreiss & al., OpenPifPaf: Composite fields for semantic keypoint detection and spatio-temporal association, in *TITS*, 2021.

2D Pose Models Finetuning

Our Framework:

- Dataset: Drive & Act.
- Metrics: AP (↑) and AR (↑).
- Finetune on **30 epochs**.
- Augmentations: scale, noise, blur.
- Specificity: Apply a **binary mask** on the joints loss **discard Feet Pose predictions.**

École des Ponts ParisTech

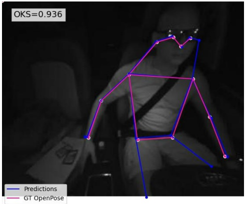
valeo.ai

2D Pose: Visual Results

(c) HR-Net: No augmentation during training

(d) HR-Net:Geometric Augmentations

(e) HR-Net: Geometric + Noise + Blurs Augmentations



(f) OpenPifPaf: Geometric augmentations (ShuffleNet-v2k16 backbone)

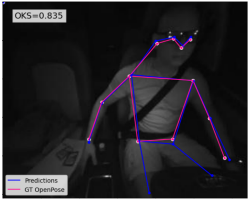


Table: Visualization of the retrained models on Drive & Act test set.

École des Ponts ParisTech

valeo.ai

2D Pose: Quantitative Results

HR Net	Input	AP	AP50	AP75	AR	AR50	AR75
No Finetuning	256 x 192	85.0	96.5	90.2	90.9	98.7	93.7
Finetuned (no aug.)	256 x 192	87.0	98.1	90.8	90.3	98.7	93.9
Finetuned (with geom. aug.)	256 x 192	90.1	**99.0**	**94.2**	**93.7**	99.4	**96.0**
Finetuned (with geom. aug. + noise + blur)	256 x 192	**90.4**	98.6	92.2	91.2	**99.5**	94.2
OpenPifPaf	**Input**	**AP**	**AP50**	**AP75**	**AR**	**AR50**	**AR75**
Finetuned (with geom. aug.)	256 x 192	84.0	93.6	87.0	88.1	93.8	90.7

Table: AP and AR on Drive & Act test set

→ HR-Net finetuned with more augmentations **outperforms** OpenPifPaf.

2D Pose Results Analysis

Pros HR Net

1. **Better scores** obtained on Drive And Act as the model's size is $2.5\times$ bigger.
2. **More keypoints** estimated.

Pros OpenPifPaf

1. **No need** of a prior detection step.
2. Inference time is **much lower** (almost a requirement for embedded systems).
3. **More stability and consistency** across consecutive frames.

- **Conclusion:** Keep working with OpenPifPaf.

École des Ponts ParisTech

valeo.ai

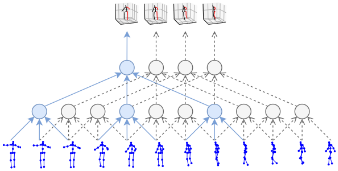
Outline

1. 2D Pose Estimation
2. 2D to 3D Pose Lifting
   - 3D Pose Lifting Model
   - 3D Pose Lifting Experiments
3. Extension of the pipeline to Face and body Pose
4. Conclusions and future work

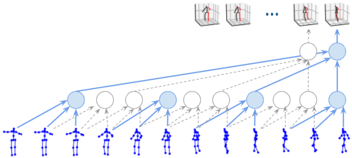
3D Pose Lifting with a CNN Model

Idea: From a sequence of 2D consecutive skeleton, predicts the 3D pose of the middle frame.

On Human3.6M[8]: **Mean Error is 37.2mm.**



(a) VideoPose3D[9] Model



(b) Causal Form of VideoPose3D

[8] Ionescu & al., Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, *TPAMI*, 2014.

[9] Pavllo & al., 3D human pose estimation in video with temporal convolutions and semi-supervised training, in *CVPR*, 2019.

3D Pose Lifting with a CNN Model

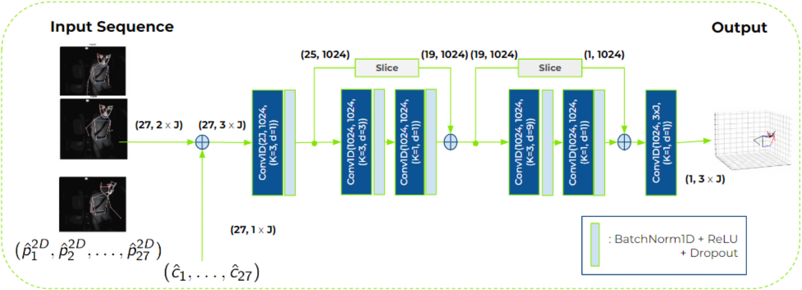


Figure: Adaptation of VideoPose3D with the addition of joints' confidence scores $\hat{c}_i$ in input.

École des Ponts ParisTech

valeo.ai

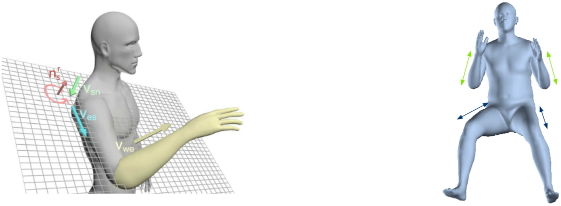
3D Pose Lifting with a CNN Model

Blocks	kernel Size length	Input frames	MPJPE(↓) (mm)	P-MPJPE(↓) (mm)	N-MPJPE(↓) (mm)	MPJVE (↓) (mm.s^{-1})
$B = 1$	$K = (3, 3)$	9	34.9±0.3	23.2±0.1	<u>27.5±0.3</u>	6.7±0.01
$B = 2$	$K = (3, 3, 3)$	27	34.6±0.5	22.8±0.1	28.0±0.3	6.63±0.03
$B = 3$	$K = (3, 3, 3, 3)$	81	33.5±0.4	22.8±0.2	27.9±0.4	6.59±0.02
$B = 4$	$K = (3, 3, 3, 3, 3)$	243	<u>33.3±0.3</u>	<u>22.6±0.1</u>	27.6±0.3	<u>6.55±0.01</u>

Table: VideoPose3D predictions on *Drive&Act* test set. with different architectures.

→ No major difference with bigger architecture.

3D Pose Lifting with a CNN Model



(a) Angles Constraint (b) Symmetry Constraint

Figure: Kinematics Constraints added

$$\mathcal{L}_{\text{sym}}(\hat{p}) = \sum_{((i,j),(k,l))\in M} \left(\|\hat{p}_i - \hat{p}_j\|_2 - \|\hat{p}_k - \hat{p}_l\|_2\right)^2 \tag{1}$$

$$\mathcal{L}_{\text{illegal}}(\hat{p}) = \exp\left(-\min(\overrightarrow{n_s^r} \cdot \overrightarrow{v_{we}}, 0)\right) \tag{2}$$

3D Pose Lifting with a CNN Model

Model	MPJPE (↓)	P-MPJPE (↓)
$\lambda_{sym} = 0.$	34.6±0.5	22.8±0.1
$\lambda_{sym} = 1.10^{-4}$	33.9±0.4	<u>22.6±0.2</u>
$\lambda_{sym} = 1.10^{-3}$	34.5±0.3	23.0±0.1
$\lambda_{sym} = 1.10^{-2}$	34.9±0.4	22.9±0.1
$\lambda_{sym} = 1.10^{-1}$	<u>33.5±0.4</u>	23.8±0.1
$\lambda_{sym} = 1.10^{0}$	50.0±0.6	43.7±0.3
$\lambda_{sym} = 1.10^{1}$	111.0±1.8	93.1±2.1
$\lambda_{sym} = 1.10^{2}$	194.4±19.9	156.9±20.4

Table: Results when training with various weighted symmetry loss.

École des Ponts ParisTech

valeo.ai

3D Pose Lifting with a CNN Model

Model	MPJPE (↓)	P-MPJPE (↓)
$\lambda_a = 0.$	34.6±0.5	<u>22.8±0.1</u>
$\lambda_a = 1.10^{-3}$	34.5±0.4	22.7±0.1
$\lambda_a = 1.10^{-2}$	34.3±0.7	22.8±0.3
$\lambda_a = 1.10^{-1}$	34.7±0.4	22.9±0.0
$\lambda_a = 1.10^{0}$	34.3±0.3	23.0±0.1
$\lambda_a = 1.10^{1}$	<u>34.3±0.4</u>	23.6±0.3
$\lambda_a = 1.10^{2}$	35.3±0.8	25.0±1.0
$\lambda_a = 1.10^{3}$	44.2±1.9	32.4±1.4

Table: Results when training with various weighted angle loss.

3D Pose Lifting Qualitative results

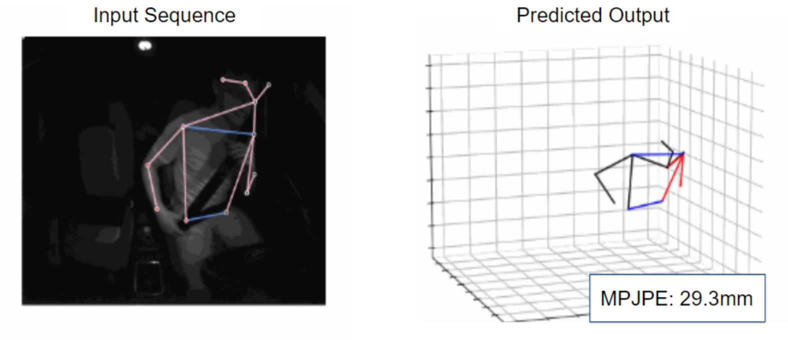


Figure: 3D Pose Prediction on Drive & Act test set.

3D Pose Lifting with a CNN Model

Conclusions:

- CNN-based VideoPose3D lifter works well with a **Mean Error around 34.0mm.**
- Study self-supervised approaches.
- Look for lighter models using transformers like **P-STMO**[10].

[10] Shan & al., P-STMO: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. in *ECCV*, 2022.

Outline

École des Ponts ParisTech

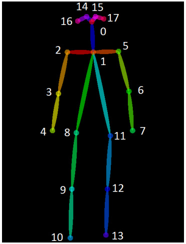
valeo.ai

Dataset Pseudo Annotation

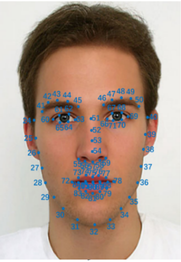
Motivation: Face Landmarks pose give better interpretability of the driver's state.

Goal: Incorporate the 3D face landmarks estimation.

Means: Use a pretrained network to estimate the 3D facial landmarks: **3DDFA v2 model**[11].



+



Refined Body Pose representation with 17 + 68 joints.[12]

[11] Guo & al., Towards fast, accurate and stable 3D dense face alignment, in *ECCV*, 2020.

[12] Jin & al., Whole-body human pose estimation in the wild, in *ECCV*, 2020.

Dataset Pseudo-Labelling

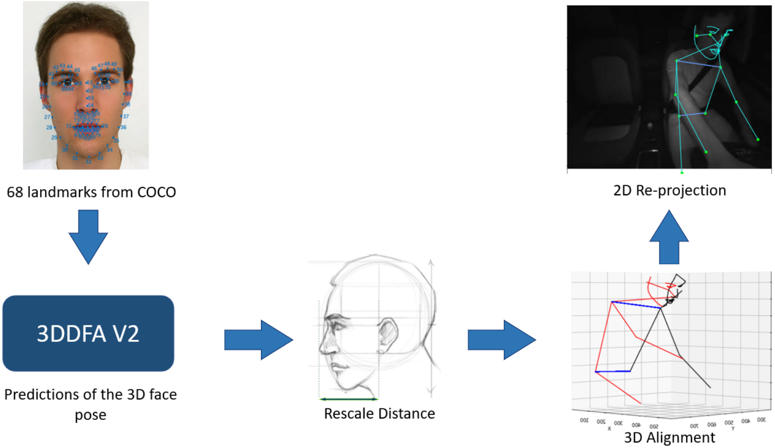


Figure: Face Alignment Protocol.

École des Ponts ParisTech

valeo.ai

Protocol Applied on Wholebody

Training Framework:

- **Input sequence:** 27 frames of $17 + 68$-joints skeletons.
- **Architecture:** 2 Blocks of Causal Convolutions with 3 dilations.
- Train on **100** epochs.
- Loss and Metric: **Mean Per Joint Error Loss**.
- Learning Rate and Batch size: 1.10^{-3} and 1024.
- Add some Dropout : **0.25**.

École des Ponts ParisTech

valeo.ai

3D Pose Lifting Results

Click for video

Occlusions Experiments

(a) VideoPose3D initial Input

(b) Added occlusions in VideoPose3D Input

Figure: Experiments on VideoPose3D's robustness, adding occlusions in the training to facilitate domain adaptation.

École des Ponts
ParisTech

valeo.ai

3D Pose Lifting Comparison Results

Click for video

Occlusions: 3D Pose Lifting Results

Model	Input frames	Occlusions ratio (%)	MPJPE(↓) (mm)	P-MPJPE(↓) (mm)	N-MPJPE(↓) (mm)	MPJVE (↓) ($mm.s^{-1}$)
VideoPose3D	27	0 %	39.4±0.8	13.4±0.2	23.8±0.2	7.28±0.02
		5 %	39.9±0.7	14.3±0.5	23.6±0.6	7.48±0.05
		10 %	40.7±1.6	14.9±0.2	24.0±1.0	7.63±0.02
		20 %	41.2±0.6	15.9±0.1	24.9±0.5	7.88±0.08
		30 %	42.5±0.7	16.3±0.1	26.8±0.2	8.07±0.05
		40 %	41.8±0.3	16.7±0.2	27.0±0.9	8.19±0.10
VideoPose3D	243	0 %	37.4±0.6	12.6±0.4	18.2±0.7	5.88±0.09
		5 %	43.1±0.3	14.2±0.3	25.0±0.8	6.86±0.08
		10 %	44.8±0.5	16.4±0.1	26.0±0.7	7.34±0.06
		20 %	43.4±0.3	16.6±0.2	26.6±0.3	7.64±0.03
		30 %	46.5±0.5	17.5±0.1	28.0±0.3	7.81 ± 0.06
		80 %	75.5±4.4	34.0±1.2	50.9±2.7	8.44±0.01

Table: Errors obtained when incorporating occlusions

Outline

1. 2D Pose Estimation
2. 2D to 3D Pose Lifting
3. Extension of the pipeline to Face and body Pose
4. Conclusions and future work
   - Discussions
   - Summary
   - Perspectives

École des Ponts ParisTech

valeo.ai

Limits of the method

- Work restricted on a single dataset.
- No real Ground Truth: Data is **pseudo-labelled** by OpenPose[13].
- Intented to minimize the errors by running each evaluation **5 times.**

[13] Cao & al., OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, in *TPAMI*, 2019.

Conclusion

Our Contributions:

- **Survey and exhaustive comparison** of Interior Monitoring datasets.
- Pseudo annotation and **3D face alignment** over Drive And Act dataset.
- **End-to-end framework** for Driver's 3D body and face landmarks pose estimation.
- Average error in 3D Pose Estimation at **34mm on average.**

Perspectives

Short term:

1. Extend the Pipeline with the **addition of 3D Hands Pseudo annotations**.
2. Smooth the pose estimation over consecutive frames.
3. Study **self-supervised methods deeper** to discard the lack of data issue.
4. Evaluate our framework on other datasets: Valeo collecting the data.

Long term:

1. **Lighten the model** to make it embeddable.
2. Activity Recognition based on sequence of 3D Pose Estimations.

École des Ponts ParisTech

valeo.ai

Thank you for your Attention !

References I

Guido Borghi, Marco Venturelli, Roberto Vezzani, and Rita Cucchiara.
Poseidon: Face-from-depth for driver pose estimation.
In *CVPR*, 2017.

Romain Guesdon, Carlos Crispim-Junior, and Laure Tougne.
Dripe: A dataset for human pose estimation in real-world driving settings.
In *ICCV*, 2021.

Jigyasa Singh Katrolia, Bruno Mirbach, Ahmed El-Sherif, Hartmut Feld, Jason Rambach, and Didier Stricker.
TICaM: A time-of-flight in-car cabin monitoring dataset.
arXiv preprint arXiv:2103.11719, 2021.

École des Ponts ParisTech

valeo.ai

References II

Okan Kopuklu, Jiapeng Zheng, Hang Xu, and Gerhard Rigoll.
Driver anomaly detection: A dataset and contrastive learning approach.
In *WACV*, 2021.

Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen.
DriveAct: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles.
In *ICCV*, 2019.

References III

Juan Diego Ortega, Neslihan Kose, Paola Cañas, Min-An Chao, Alexander Unnervik, Marcos Nieto, Oihana Otaegui, and Luis Salgado.
Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis.
In *ECCV*, 2020.

Mohamed Selim, Ahmet Firintepe, Alain Pagani, and Didier Stricker.
AutoPOSE: Large-scale automotive driver head pose and gaze dataset with deep head orientation baseline.
In *VISIGRAPP (4: VISAPP)*, 2020.

3D Pose Lifting with CNN model: Semi-Supervised Approach

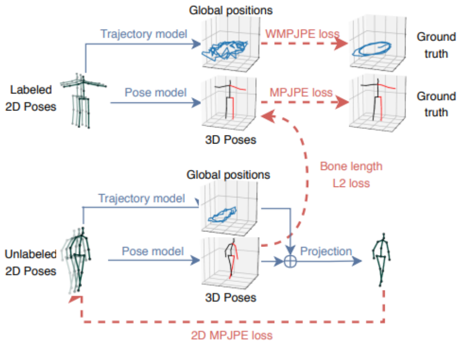


Figure: VideoPose3D Semi-supervised approach[14]

[14] Pavllo & al., 3D human pose estimation in video with temporal convolutions and semi-supervised training, in *CVPR*, 2019.

3D Pose Lifting with Transformer-based model

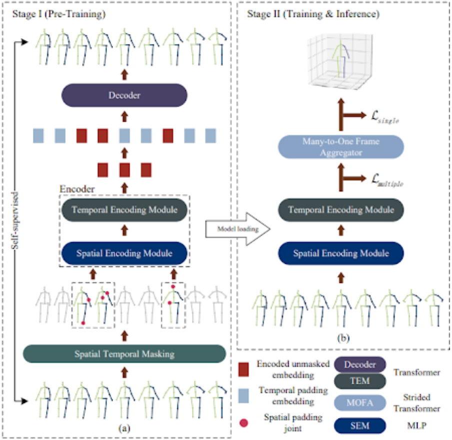


Figure: P-STMO model[15]

[15]Shan & al., P-STMO: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. in *ECCV*, 2022.