Active Transfer Learning for Efficient Video-Specific Human Pose Estimation

Hiromu Taketsugu and Norimichi Ukita (IIMLab., TTI-J)

JAN 4-8 WACV 2024 WAIKOLOA HAWAII

Introduction | Human Pose (HP) Estimation

Motivation: HP analysis

Surveillance | Healthcare | Sports

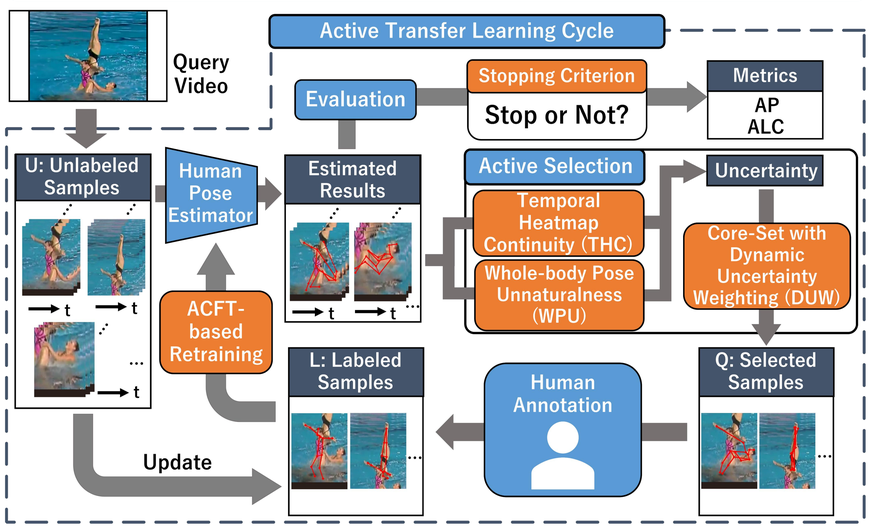
Issue: Domain gap between train/test

Dataset	Train	Test
PoseTrack21 [1]	**98.95**	**75.50**
JRDB-Pose [2]	**95.31**	**39.50**

Solution: Video-specific HP estimation

- **Active Learning (AL):** Human Annotation
- **Transfer Learning (TL):** Fine-tune the model

How to select? How to retrain? How to stop?



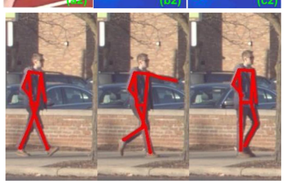
Related Work | Methods for Efficient AL

- **HP Estimation × AL:**

MPE [3]: Quantifies uncertainty from local peaks

[3]

TPC [4]: Quantifies uncertainty by temporal change



[4]

✗ **Discarding information**

✗ **Only for keypoint-level** ✗ **Selection bias**

- **Active Transfer Learning:**

ACFT [5]: Efficiently adopt CNN to specific data domain

✗ **Only for classification**

- **Stopping Criterion:**

Min-error [6]

$$\frac{1}{|Q|}\sum_{x\in Q} OKS(x) > \theta$$

✗ **Premature stop**

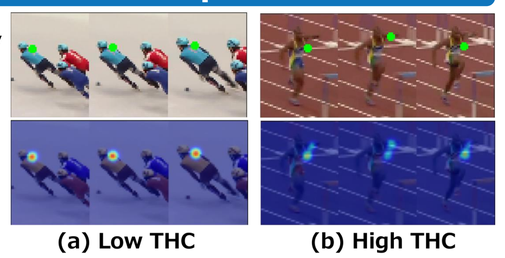
More details | Code, Project page, Preprint

Proposed Method | Video-Specific Active Transfer Learning (ATL)

How to select informative samples?

- **THC: Keypoint-level** uncertainty

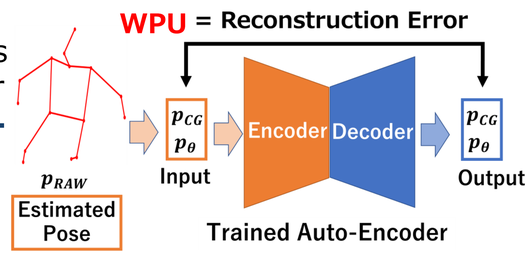
Calculate distance of temporally adjacent heatmaps by SAE
Uncertainty 👉 Inconsistent heatmap



(a) Low THC (b) High THC

- **WPU: Pose-level** uncertainty

Train Auto-Encoder with GT poses
It can't reconstruct unnatural poses
Uncertainty 👉 Reconstruction error



- **DUW:** Avoid **selection bias**

Select uncertain & diverse samples by weighting Core-Set [7] sampling.

$$(\text{New sample}) = \underset{i\in U}{\operatorname{argmax}}\{\min_{j\in L}\{(1-G_C)\times\Delta(x_i,x_j)\} + G_C \times \lambda C(x_i)\}$$

How to retrain?

- **ACFT-based Retraining:**

Adopt ACFT to regression task (HP estimation).

Reduce redundancy by selecting samples to retrain:

$$\mathrm{R} = \{x | OKS(x) < \theta + m\}$$

How to stop?

- **Novel stopping criterion:**

Guarantee ***all*** labeled samples reach desired accuracy.

Terminate ATL if:

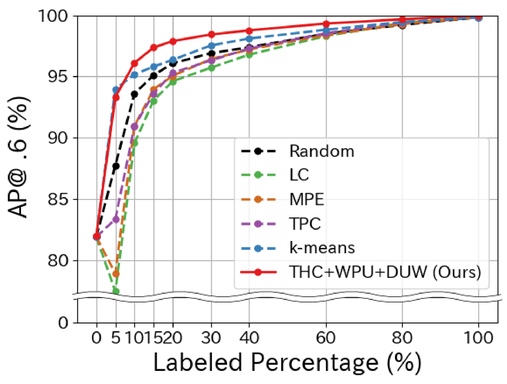
$$\forall x \in \{Q \cup L\}, OKS(x) > \theta$$

Experiments

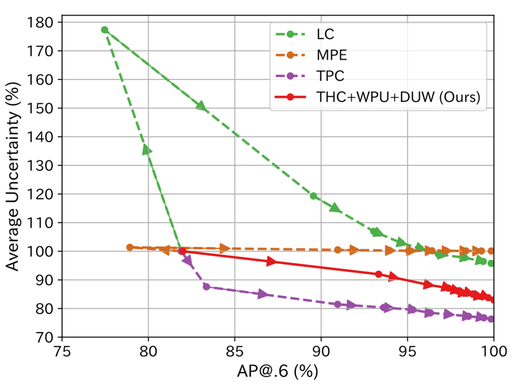
- **Dataset:** PoseTrack21 [1], JRDB-Pose [2]
- **Metric:** Area under the Learning Curve (ALC), Average Precision (AP)

Quantitative results

Criterion	PoseTrack21	JRDB-Pose
Random	96.91	95.42
LC	95.74	92.67
MPE [3]	96.11	95.76
TPC [4]	96.40	93.76
k-means [8]	**97.65**	**96.41**
Ours	**98.21**	**96.52**



Detailed analysis

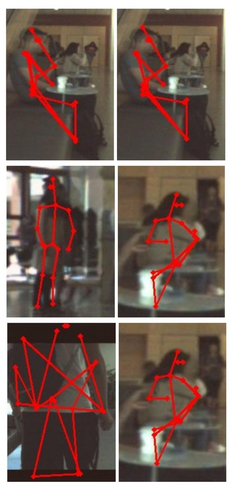


Ablation study

Criterion	AP@0.6 (%) 5%	AP@0.6 (%) 20%	ALC (%)
Core-Set [7]	93.18	97.62	98.12
THC	82.59	92.86	95.45
WPU	85.56	94.74	96.45
THC+WPU	84.82	95.17	96.51
THC+DUW	93.12	97.70	**98.19**
WPU+DUW	**93.19**	**97.87**	98.17
Fixed G_C	93.02	97.68	98.14
Ours (full)	**93.35**	**97.90**	**98.21**

Qualitative results

- **MPE** [3]: Selection bias
- **Core-Set** [7]: Easy samples
- **Ours**: Uncertain and diverse



Stopping Criterion

SC	θ	AP@θ (%)	Stopped (%)	Actual (%)
SC_{Min} [6]	0.5	**96.86**	10.04	36.90
	0.6	**96.19**	10.85	40.26
	0.7	**95.75**	12.55	46.55
	0.8	**95.58**	16.94	56.54
SC_{All} (Ours)	0.5	**99.25**	29.61	36.90
	0.6	**99.55**	33.38	40.26
	0.7	**99.60**	39.61	46.55
	0.8	**99.50**	49.46	56.54

Our method enables efficient video-specific HP estimation through ATL 👍

Future Work

- Using video-based HP estimators
- Incorporating semi-supervised learning

Acknowledgement

- This work was supported by the NEC C&C Foundation Travel Grant 🙏

Reference

[1] Doering et al., CVPR2022 [2] Vendrow et al., CVPR2023 [3] Liu and Ferrari, ICCV2017 [4] Mori et al., IWAIT2022
[5] Zhou et al., Med. Image Anal., 2021 [6] Zhu et al., IJCNLP2008 [7] Zhdanov, arXiv, 2019 [8] Sener and Savarese, ICLR2018

Background: https://unsplash.com/photos/a-lake-surrounded-by-palm-trees-under-a-blue-sky-OAiTRB8zQYI