UNIVERSITÄT
KOBLENZ · LANDAU

Video Based Human Action Recognition Using Deep Learning

Why This Thesis

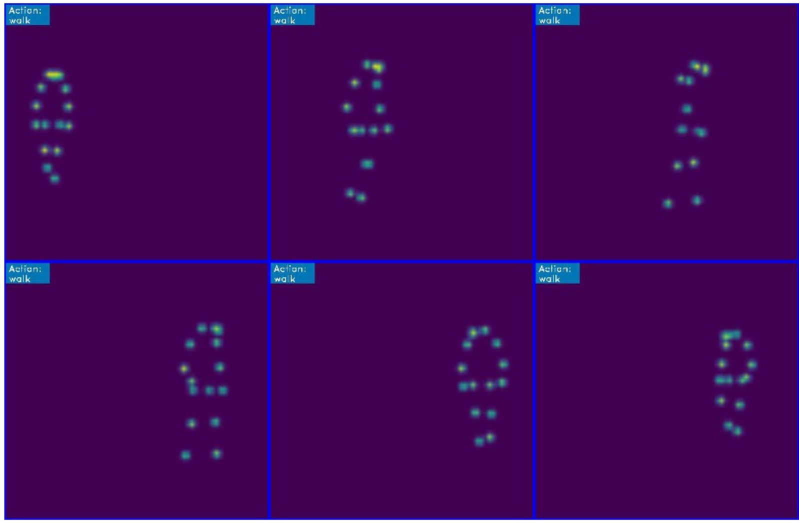
- Integrate machine learning to university level dance education
- Investigate the use of deep action recognition models
- Investigate the provided BAST dataset

Bewegungs Analyse Skalen UND Test

- Based on the Laban analysis (body, effort, shape, space)
- Investigates the relation between movements & mental state
- Nine basic movements evaluated on specific dimensions

Human Action Recognition

- Sprinting: crouch (start) → run on track (middle) → finish (end line)
- Input modality: RGB, Optical Flow, 2D poses
- Goal: recognize BAST movements



2D poses made of 17 keypoints

Datasets

- 98 videos, partly annotated; using 10s clips with 5s sliding window
- Formed two datasets for base and evaluation annotations respectively
- Insufficient data & imbalance data for bast-eval

Table 1: Number of clips for the *bast-base* and *bast-eval* datasets

Dataset	#Clips
bast-base	3894
bast-eval	4112

Experiments

- Best Classifier for Base Annotations
- Best Classifier for Evaluation Annotations
- Robustness of Models

Experiments

- Transfer Learning
- Background Influence

Best Classifier for Base Annotations

- 2D Conv Nets either underfit or overfit
- 3D Conv Nets perform an excellent job
- PoseC3D also very fast to train and low complexity (params & flops)

Model	Caveat	Top1 Acc	Top2 Acc	Top2 Acc (val)	Mean Cls Acc
I3D	*baseline*	87.2%	96.3%	92.6%	87.1%
	no dropout	20.9%	25.3%	86.4%	18.4%
SlowOnly	*omni-pr*	92.1%	**97.31%**	95.1%	91.7%
SlowFast	*baseline*	90.6%	96%	95%	90%
CSN	*baseline*	-	-	30%	-
PoseC3D	*gym-pr*	88.9%	95.4%	94%	88.6%
	gym-0.7d	90.61%	95.5%	97.1%	90%
	ntu60-pr	89.25%	95.5%	96.3%	89.2%
	ntu120-0.8d	91.5%	96.25%	**97.8%**	91.1%
	ntu120-0.8d-54x1x1	87.0%	93.9%	96.9%	85.7%
	ntu120-0.8d-64x1x1	88.9%	96.1%	97.6%	91%
	kinetics-ucf	**92.32%**	**97.27%**	97.7%	**92.52%**
	kinetics-0.7d-32x1x1	90.44%	96.93%	97.4%	89.6%

Best Classifier for Eval Annotations

- PoseC3D provides pretty decent results
- 64x1x1 dense sampling strategy for fine-grained actions
- Half of annotations classified perfectly

Model	Caveat	Top3 Acc (test)	Top3 Acc (val)	Mean Class Acc
I3D	-	59.5%	54.7%	25.4%
	0.4d	27%	63%	5.4%
	0.65d	22%	58%	9.8%
	0.6d-48x3x1	66%	59%	21%
SlowOnly	*bb-pr-8x8x1*	60.3%	64.4%	24.6%
	bb-pr-0.6d-16x8x1	71%	67.6%	25.64%
PoseC3D	*gym-pr*	69%	69.4%	22%
	gym-bb-pr-0.65d	72.2%	72.3%	32.5%
	ntu120-pr-0.7d	71.2%	**74.9%**	32.2%
	ntu120-pr-0.8d	75.7%	68.5%	23.6%
	bb-pr-64x1x1-0.6d	**77.39%**	72.68%	**33.19%**

Robustness of Models

- Nine bast-base movements as general categories
- No evaluation dataset

- Derive heuristics from bast-avatar and explain model's performance

Figure 4.5: Emulation of *water* using body movements from the *bast-avatar* test dataset

Water Heuristics

- Emulate wave patterns with arms
- Hand movements point to horizontal direction
- Perform fish-like movements

Fire Heuristics

- Emulating something going up in the air with hand movements
- Imitate explosions by jumping

Air Heuristics

- Move around the room a lot
- Move hands a lot
- Rotate to simulate a whirl

Earth Heuristics

- Fall to the ground and move on all fours
- Some people contracted on the ground
- Some hit the ground with legs

UNIVERSITÄT KOBLENZ · LANDAU

Results

Robustness of Models

- Test the kinesphere evaluation of contract/expand
- Validation dataset contains complex, dance-like kinesphere
- Evaluations: narrow, middle, wide

Figure 4.4: Kinesphere movement from the *kinesphere* test dataset

Results

- The four related movements: *expand-narrow; expand-wide; expand-wide-more; expand-narrow-more* entirely missing
- The model outputs movements not related to the kinesphere
- No subtle difference between predictions for *narrow-kinesphere; middle kinesphere;* and *wide-kinesphere*

Table 11: *kinesphere* domain-test dataset analysis

Ground truth	Model's prediction	Count
narrow kinesphere	*jump-time-long*	13
	stamp-body-isolated	12
	jump-emphasis-upward	9
middle kinesphere	*jump-time-long*	7
	con-expand-no-emphasis	7
	walk-straight-more	6
wide kinesphere	*stamp-strength-none*	9
	stamp-body-isolated	6
	jump-time-long	6

Transfer Learning

- Improvement of learning in a new task through the transfer of knowledge from an already learned task

- Ameliorates the insufficent samples and imbalanced classes problem for tasks that have small datasets

Training From Scratch vs Transfer Learning

Table 12: Training from scratch vs. transfer learning

Model	Task type	Transfer learning	No transfer learning
I3D	*bast-base*	96.3%	87%
	bast-eval	60%	51%
SlowOnly	*bast-base*	97.31%	90.78%
	bast-eval	71%	29.31%
SlowFast	*bast-base*	96%	89.8%
PoseC3D	*bast-base*	97.27%	95.9%
	bast-eval	77.39%	72.19%

Transfer Learning with Bast Base for Bast Eval

Table 14: Benchmark dataset vs *bast-base* dataset pre-training for the *bast-eval* task

Model	**Benchmark**	**Bast-base**
I3D	59.5%	62.3%
SlowOnly	60.3%	70%
PoseC3D	75.7%	77.39%

Transfer learning for Bast Eval

- Also solves the class imbalance problem to a certain degree

- Nevertheless, class imbalance remains a problem

- Best benchmark dataset for pre-training: Kinetics & Omni-Sourced

Background Influence

- Background plays an important role in the model's prediction
- Ideally the model should only focus on the person inside the frames
- "Dancing ballet" vs "jogging"

GradCam Analysis

UNIVERSITÄT KOBLENZ · LANDAU

GradCam Analysis on Models Without Background

Background Influence

- Stripping background hurts the model's performance
- Models are less robust on testing sets
- Solution: Use 2D poses as input stream

Conclusions

- Perfect classification for base annotations; good results for evaluation
- Base classifier possibly extendable; eval classifier too specific for BAST
- Transfer learning yields robust models; background is crucial to model

UNIVERSITÄT KOBLENZ · LANDAU

Appendix

BAST Evaluation

- Floor Pattern: (1) rather straight (2) rather curved (3) curved
- Emphasis: (1) upwards (2) forward
- Time in air: (1) long (2) short
- Body involvement: (1) isolated (2) whole body
- Strength: (1) no strength (2) little (3) max
- Kinesphere: (1) narrow (2) medium (3) wide
- Emphasis: (1) contracting (2) expanding (3) none
- Balance: (1) unstable (2) rather stable (3) stable
- Flow: (1) very bound (2) bound (free) (4) very free)
- Acceleration: (1) yes (2) no
- Falling-flow: (1) lying down (2) free
- End-position: (1) sitting (2) lying

Training With Various Benchmark Datasets

- Top benchmark datsets: Kinetics 400 & Omni-Sourced

Table 13: Training with various benchmarks for both tasks

Model	Benchmark	Bast-Base	Bast-Eval
SlowOnly	Omni-Sourced	97.31%	29.3%
	Kinetics400	24.07%	46.4%
TIN	Sth-Sth-V2	19.7%	-
	Kinetics400	28.45%	-
PoseC3D	Gym	96.93%	74%
	Ntu-60	95.5%	-
	Ntu-120	96.25%	75.7%
	Kinetics-hmdb	97.27%	72.19%
	Kinetics-ufc	97.42%	71.36%

Class Imbalance Problem of Bast-Eval

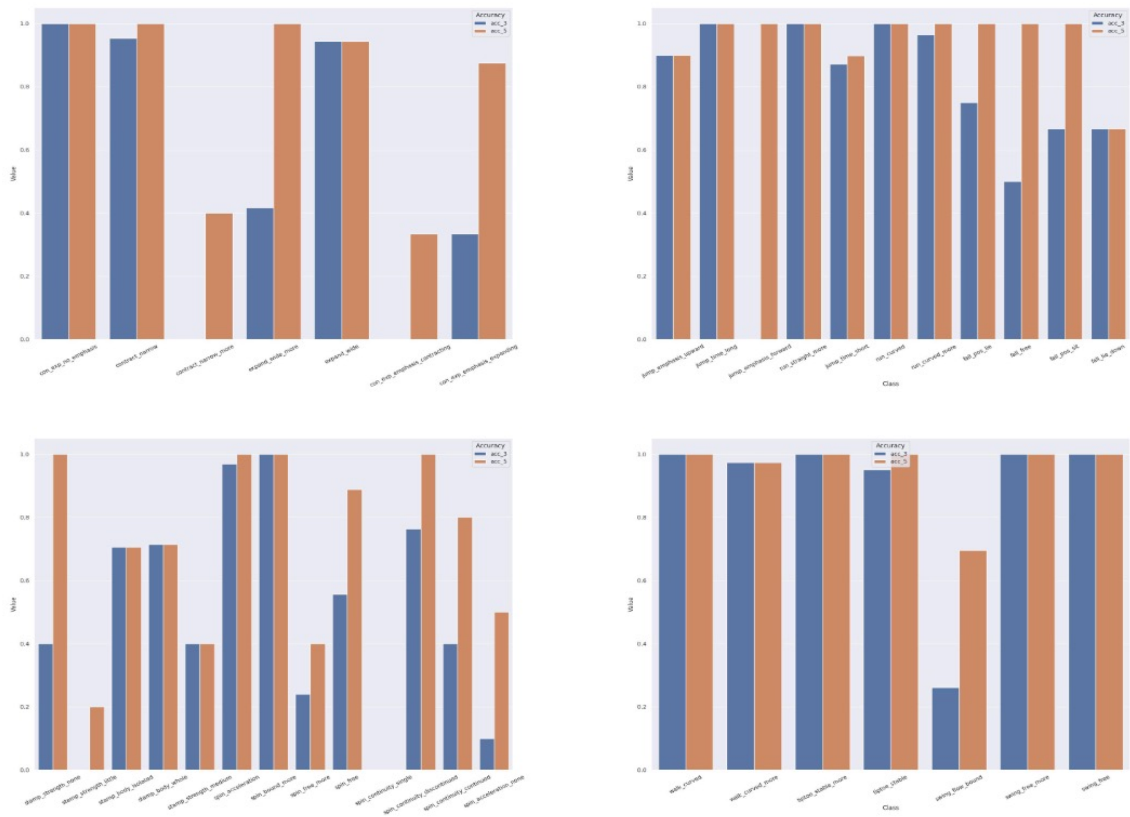


Figure 6.7: Accuracy for each annotation of the *bast-eval* task for the PoseC3D model pre-trained on the *ntu-120* dataset

Class Imbalance Problem of Bast-Eval

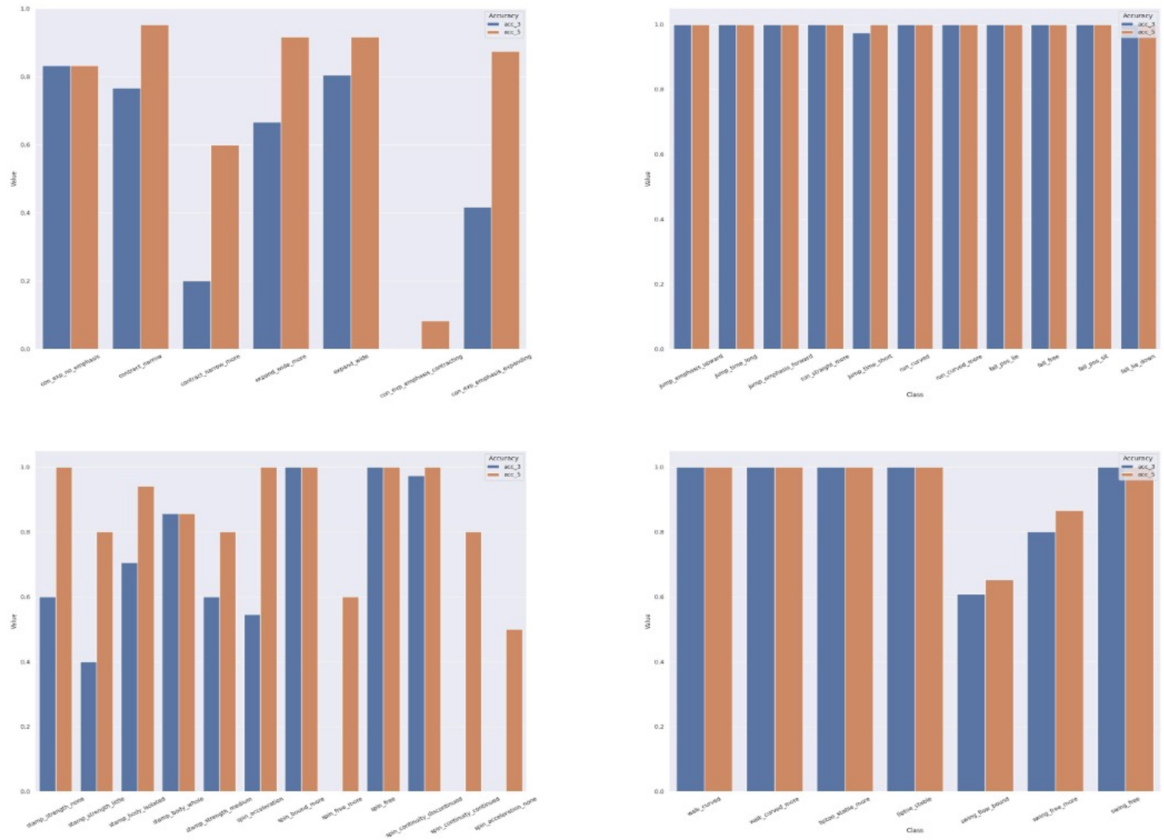


Figure 6.5: Accuracy for each annotation of the *bast-eval* task for the PoseC3D model pre-trained on the *bast-base* dataset

UNIVERSITÄT KOBLENZ · LANDAU

Class Imbalance Problem of Bast-Eval

- the *kinesphere* evaluation of *contract* is imbalanced as the evaluation *contract-narrow* has more than double the amount of annotations that *contract-narrow-more* has. In Figure 6.7 it is observed that the model is unable to correctly classify this evaluation. The *top-3* accuracy for the *contract-narrow-more* annotation is 0, while the classification of *contract-narrow* is almost perfect. In contrast, in Figure 6.5 we can see that the model has started to recognize the *contract-narrow-more* evaluation.

- the *kinesphere* evaluation of *expand* is imbalanced as the evaluation *expand-wide* has more than triple the amount of annotations that *expand-wide-more* has. The accuracy of *expand-wide-more* is roughly 40% for the PoseC3D model trained on *ntu-120* but with a *bast-base* pre-training this accuracy rises to almost 70%. Thus the latter model is able to recognize both the classes properly even though they are severely imbalanced.

- the *emphasis* evaluation of *jump* is in particular severely imbalanced because the evaluation *jump-emphasis-upward* has as much as eight times more samples than its counterpart *jump-emphasis-forward*. Notwithstanding this, the model pre-trained on *bast-base* is able to perfectly classify both of them when taking into account the *top-3* accuracy. However, the model not pre-trained on *bast-base* cannot even classify one sample correctly when considering the *top-3* accuracy for the *jump-emphasis-forward* evaluation. This is understandable given the severe imbalance

Class Imbalance Problem of Bast-Eval

but the fact that a pre-training with *bast-base* solves this problem perfectly really serves to prove the point of this section.

- the *strength* evaluatuion of *stamp* is slightly imbalanced because *stam-strength-medium* has almost the same amount of annotations as *stamp-strength-none*, and *stamp-strength-little* taken together. For the model not pre-trained on *bast-eval*, the *top-3* accuracy for *stamp-strength-none* is 40% and for *stamp-strength-little* is 0%. The model pre-trained on *bast-base* on the other hand has a confidence of 60% and 40% respectively.