Temporal Smoothing in 2D Human Pose Estimation for Bouldering

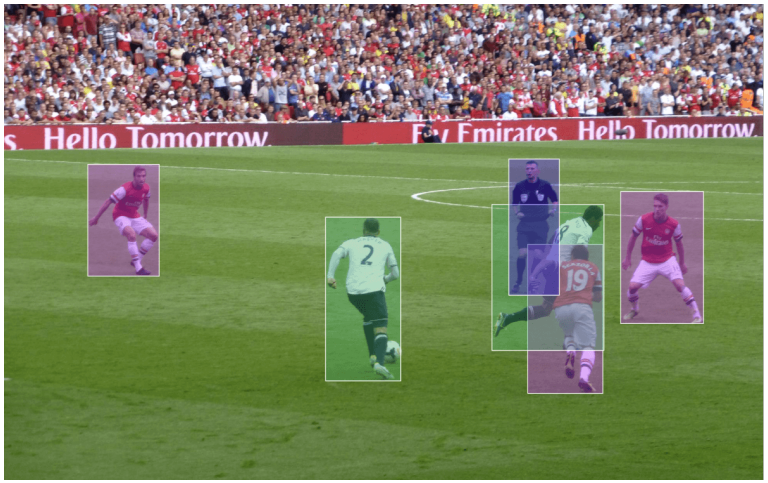
André Oskar Andersen
wpr684

Institution of Computer Science, University of Copenhagen

2023

Introduction

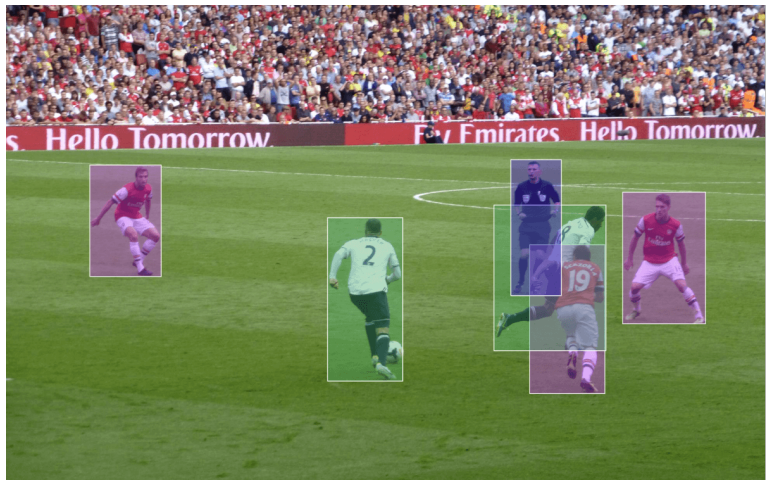
- Increased usage of video analysis in sports.
  - Help referee
  - Improve techniques



https://www.superannotate.com/blog/computer-vision-in-sports

Introduction

- Increased usage of video analysis in sports.
- Often requires the position of the players.
  - Already developed for popular sports.
  - Missing for the less popular sports.



https://www.superannotate.com/blog/computer-vision-in-sports

Introduction

- Increased usage of video analysis in sports.
- Often requires the position of the players.
- Problems with the data
  - Methods require a lot of data
  - Unusual poses/movements

https://climbalong.com/lab

Introduction

- Increased usage of video analysis in sports.
- Often requires the position of the players.
- Problems with the data
- ClimbAlong at NorthTech ApS
  - Frame-independent pose-detector for bouldering - suboptimal results

https://climbalong.com/lab

Introduction

- Increased usage of video analysis in sports.
- Often requires the position of the players.
- Problems with the data
- ClimbAlong at NorthTech ApS
  - Frame-independent pose-detector for bouldering - suboptimal results
  - Proposition: Incorporate temporal information

https://climbalong.com/lab

Introduction

- Aim: extend the ClimbAlong pose-detector to use temporal information.

The Models

- Generally, three approaches
  1. Convolutional layer
  2. Recurrent neural network (RNN)
  3. Transformer

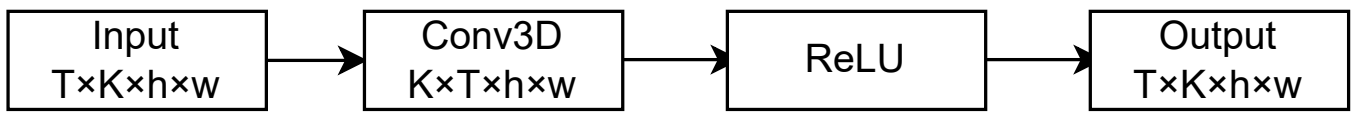
The Models

- Generally, three approaches
  1. Convolutional layer
  2. Recurrent neural network (RNN)
  3. Transformer
- One of each approach

The Models

Convolutional layer

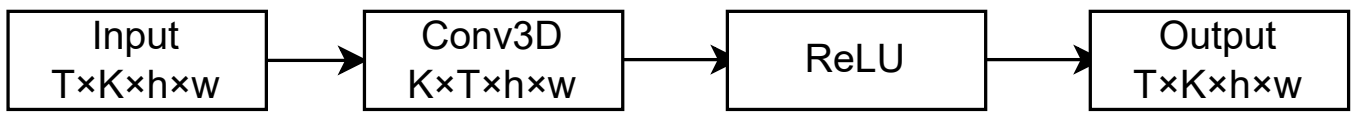
- Name: 3DConv



The Models

Convolutional layer

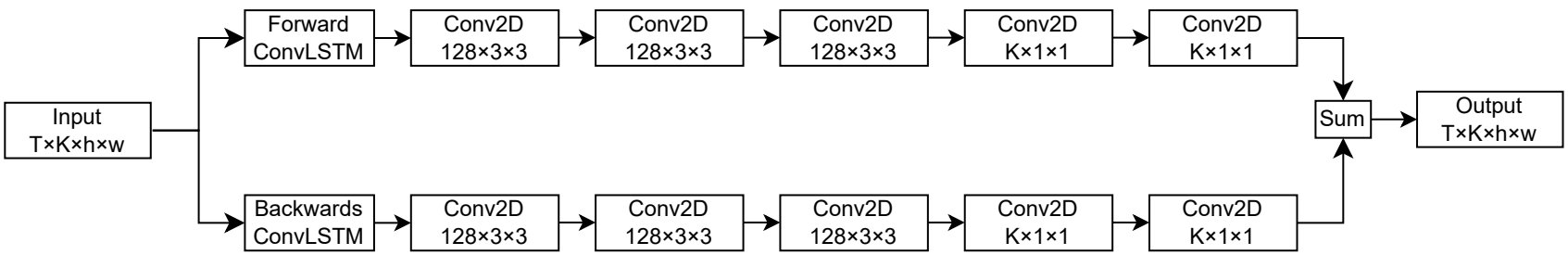
- Name: 3DConv
- 3-dimensional conv. layer + ReLU



The Models

RNN-based 1:

- Name: bi-ConvLSTM Model S
- Adaptation of Unipose-LSTM by Artacho and Savakis
- Bidirectional convolutional LSTM + 2D-conv. layers and ReLUs
- Processing directions summed together



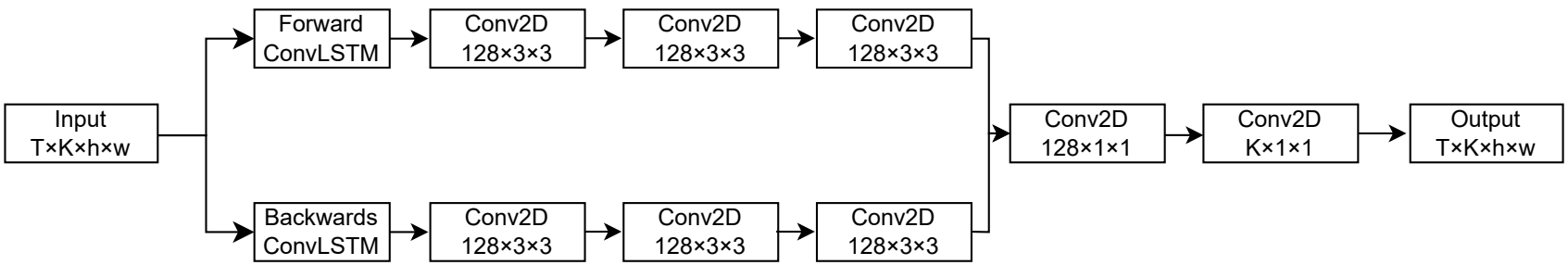
---

https://openaccess.thecvf.com/content_CVPR_2020/papers/Artacho_UniPose_Unified_Human_Pose_Estimation_in_Single_Images_and_Videos_CVPR_2020_paper.pdf

The Models

RNN-based 2:

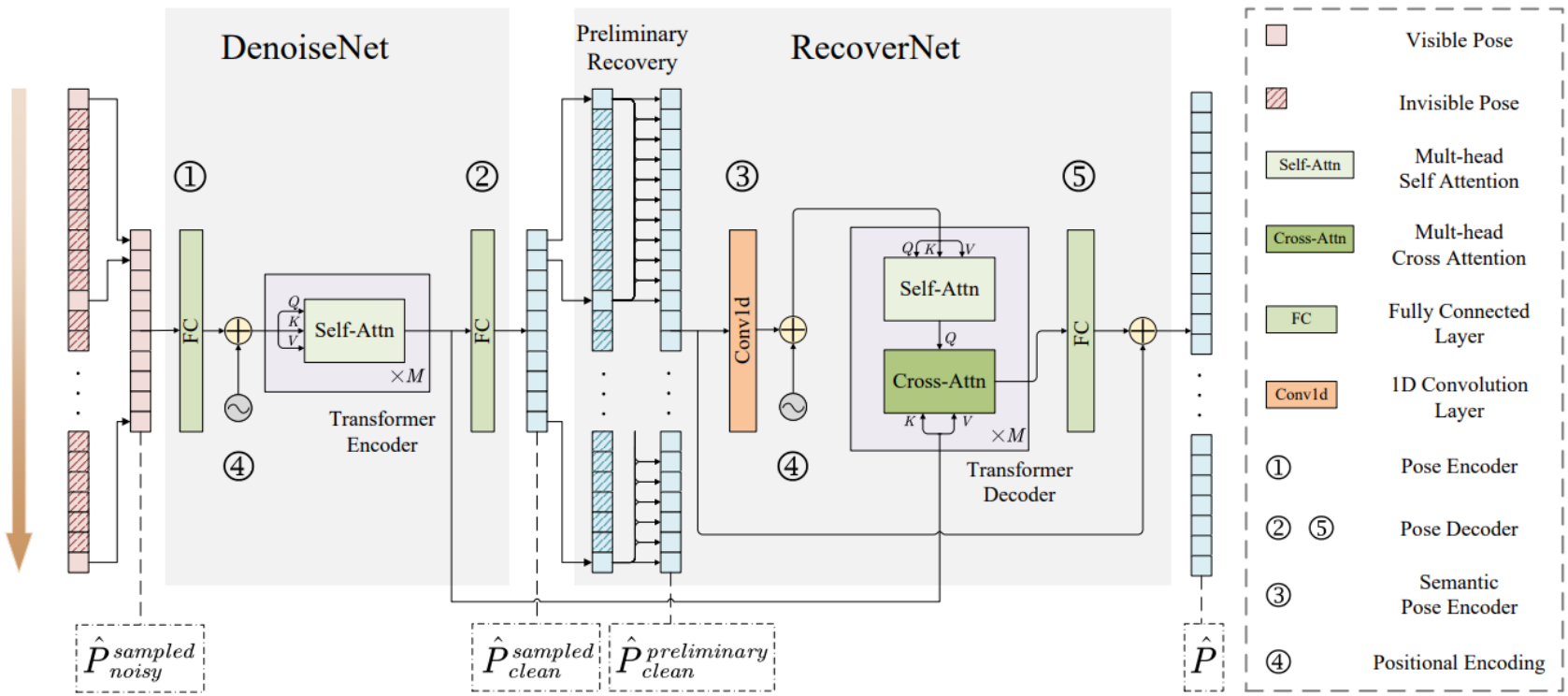
- bi-ConvLSTM Model C
- Problem: Prioritization of processing direction
- Solution: Using convolution
- Otherwise, very similar to bi-ConvLSTM Model S



The Models

DeciWatch by Zeng *Et al*.

- Transformer-based
- Samples every *n*th frame
- DenoiseNet + RecoverNet



https://arxiv.org/pdf/2203.08713.pdf

The Data

ClimbAlong

- Fully annoated videos of climbers

https://climbalong.com/lab

The Data

ClimbAlong

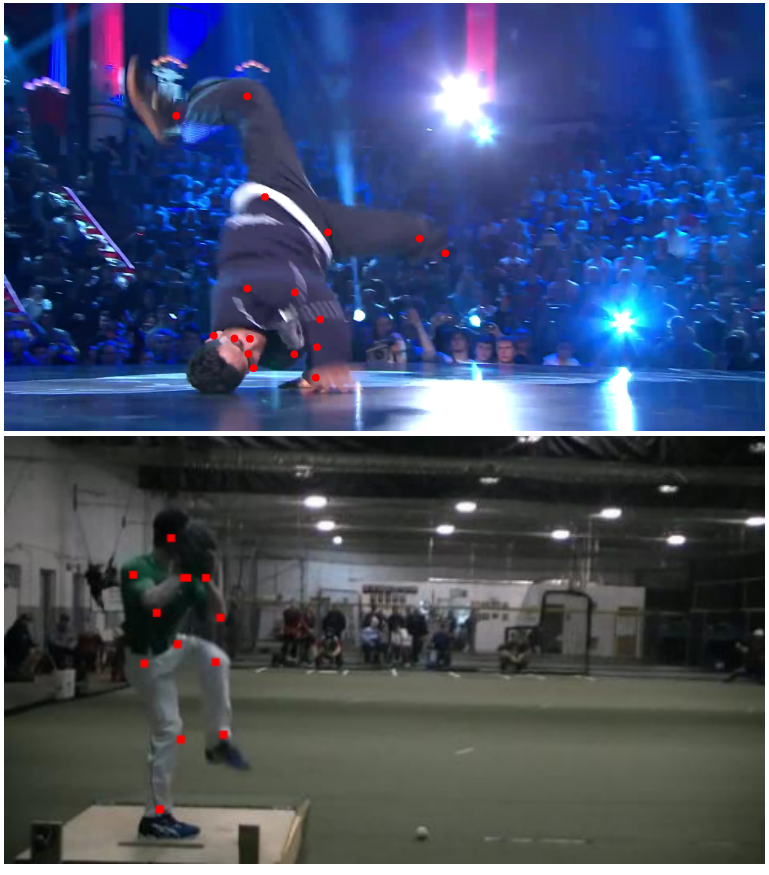
- Fully annoated videos of climbers
- Problem: very small dataset

https://climbalong.com/lab

The Data

ClimbAlong

- Fully annoated videos of climbers
- Problem: Very small dataset
- Solution: pretrain on related datasets and finetune on ClimbAlong
  - BRACE
  - Penn Action



Data Configuration

- $s = 5$ frames, as suggested by Artacho *Et al.*

---

https://arxiv.org/pdf/2001.08095.pdf

Data Configuration

- $s = 5$ frames, as suggested by Artacho *Et al.*
  1. Less memory

---

https://arxiv.org/pdf/2001.08095.pdf

Data Configuration

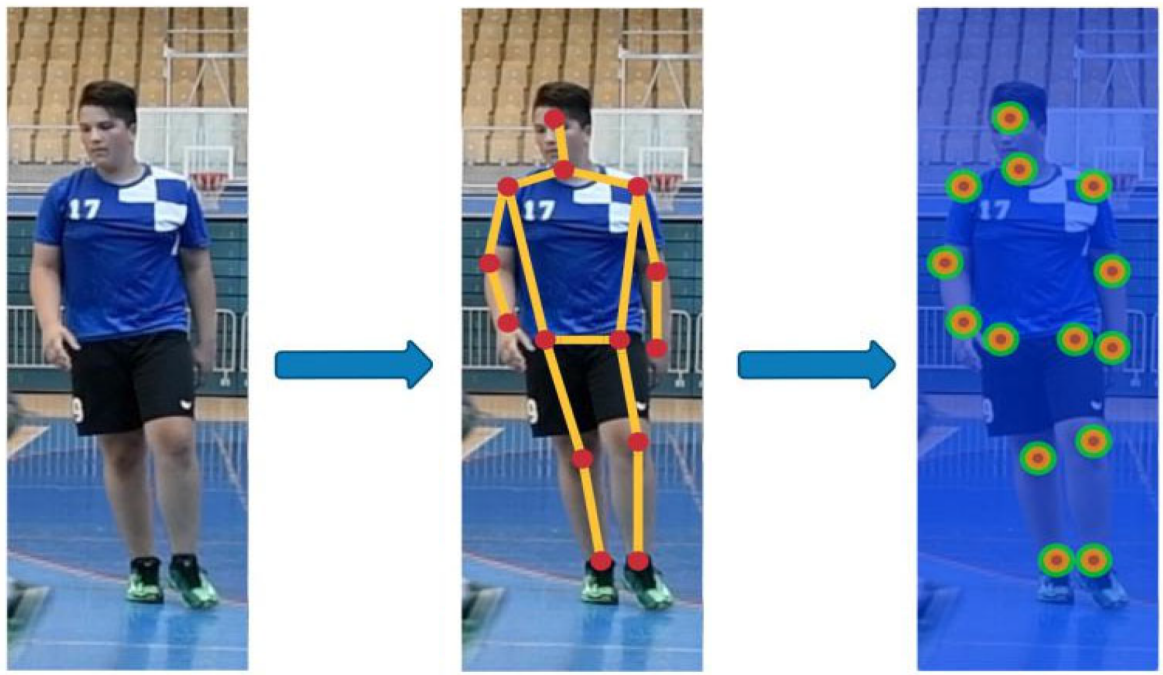
- $s = 5$ frames, as suggested by Artacho *Et al.*
  1. Less memory
  2. Model-hyperparameter based on video length

---

https://arxiv.org/pdf/2001.08095.pdf

Data Configuration

- $s = 5$ frames, as suggested by Artacho *Et al.*
  1. Less memory
  2. Model-hyperparameter based on video length
- Creation of heatmaps



---

https://www.mdpi.com/2313-433X/8/11/308

https://arxiv.org/pdf/2001.08095.pdf

Pretraining

Procedure

- Not training already-developed pose-detector
  - Different input images = Unrepresentative pose-detector predictions

Pretraining

Procedure

- Not training already-developed pose-detector
  - Different input images = Unrepresentative pose-detector predictions
- Instead, simulate pose-detector output by shifting input keypoints = learn to denoise

Pretraining

Finding optimal setting of models

- Three experiments
  1. Various smearing standard deviations: pose-detector does not use fixed standard deviation

Pretraining

Finding optimal setting of models

- Three experiments
  1. Various smearing standard deviations: pose-detector does not use fixed standard deviation
  2. Fixed smearing standard deviation: removed some of randomness from experiment 1

Pretraining

Finding optimal setting of models

- Three experiments
  1. Various smearing standard deviations: pose-detector does not use fixed standard deviation
  2. Fixed smearing standard deviation: removed some of randomness from experiment 1
  3. Various smearing standard deviations + decreased frame rate: increased context

Pretraining

Finding optimal setting of models

- Three experiments
  1. Various smearing standard deviations: pose-detector does not use fixed standard deviation
  2. Fixed smearing standard deviation: removed some of randomness from experiment 1
  3. Various smearing standard deviations + decreased frame rate: increased context
- Two different shifting-scalars

Finetuning

Freezing already-developed pose-detector

1. Quicker fitting
2. Greater understanding of results

Accuracy metric

Percentage of correct keypoints (PCK)

- Correct if $dist(pred, gt) \leq d$
- PCK@0.05, PCK@0.1 and PCK@0.2

Test results

Only minor difference:

- Shifting-scalar: shifting-scalar $s = 1$ better simulates the pose-detector
- Translation + scaling vs only translation: standard deviations of peaks in ClimbAlong data not changing as much

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	49.7	52.3	53.1	95.7	95.7	95.8	99.2	99.3	99.3
DeciWatch	76.6	76.7	68.1	94.4	94.3	87.3	99.2	99.2	96.1
bi-ConvLSTM - Model S	37.8	34.9	39.0	91.8	92.1	92.2	99.4	99.7	99.2
bi-ConvLSTM - Model C	35.9	39.0	38.5	93.1	93.6	92.6	99.8	99.7	99.7

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	46.5	51.6	47.3	95.5	95.5	95.8	99.2	99.3	99.2
DeciWatch	76.0	75.9	36.8	94.2	94.2	74.9	99.2	99.2	92.8
bi-ConvLSTM - Model S	38.8	37.4	35.9	92.7	92.1	91.2	99.4	99.5	99.3
bi-ConvLSTM - Model C	39.2	39.5	37.1	92.5	92.9	92.6	99.6	99.3	99.6

Test results

Decreased frame rate:

- 3DConv: benefit
- DeciWatch: drawback
- bi-ConvLSTM: benefit for less noise

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	49.7	52.3	53.1	95.7	95.7	95.8	99.2	99.3	99.3
DeciWatch	76.6	76.7	68.1	94.4	94.3	87.3	99.2	99.2	96.1
bi-ConvLSTM - Model S	37.8	34.9	39.0	91.8	92.1	92.2	99.4	99.7	99.2
bi-ConvLSTM - Model C	35.9	39.0	38.5	93.1	93.6	92.6	99.8	99.7	99.7

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	46.5	51.6	47.3	95.5	95.5	95.8	99.2	99.3	99.2
DeciWatch	76.0	75.9	36.8	94.2	94.2	74.9	99.2	99.2	92.8
bi-ConvLSTM - Model S	38.8	37.4	35.9	92.7	92.1	91.2	99.4	99.5	99.3
bi-ConvLSTM - Model C	39.2	39.5	37.1	92.5	92.9	92.6	99.6	99.3	99.6

Test results

bi-ConvLSTM: Model S vs Model C:

- Not as a big of a concern

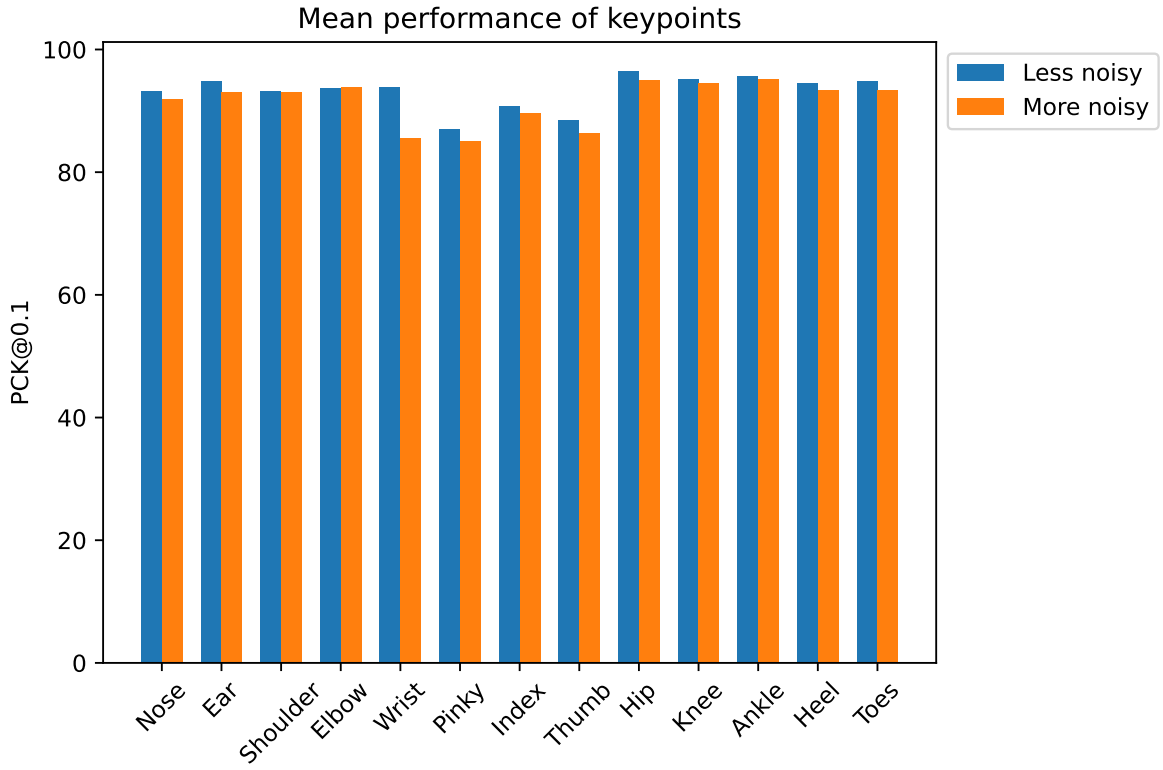
Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	49.7	52.3	53.1	95.7	95.7	95.8	99.2	99.3	99.3
DeciWatch	76.6	76.7	68.1	94.4	94.3	87.3	99.2	99.2	96.1
bi-ConvLSTM - Model S	37.8	34.9	39.0	91.8	92.1	92.2	99.4	99.7	99.2
bi-ConvLSTM - Model C	35.9	39.0	38.5	93.1	93.6	92.6	99.8	99.7	99.7

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	46.5	51.6	47.3	95.5	95.5	95.8	99.2	99.3	99.2
DeciWatch	76.0	75.9	36.8	94.2	94.2	74.9	99.2	99.2	92.8
bi-ConvLSTM - Model S	38.8	37.4	35.9	92.7	92.1	91.2	99.4	99.5	99.3
bi-ConvLSTM - Model C	39.2	39.5	37.1	92.5	92.9	92.6	99.6	99.3	99.6

Test results

Worst performing keypoints:

- Pinkies, index fingers and thumbs
- Not included during pretraining (minor effect - heels and toes)
- A lot of movement



Which model is the optimal choice?

- Greatest testing accuracy: DeciWatch shifting-scalar $s = 1$, full frame rate
- Greatest rough estimation: bi-ConvLSTM Model C, shifting-scalar $s = 1$, experiment 1
- Speed and memory: 3DConv, shfting-scalar $s = 1$, experiment 3

General reflections

Mistakes we have made

- Pretraining
  - Should have estimated parameters of data

General reflections

Mistakes we have made

- Pretraining
  - Should have estimated parameters of data
  - Same person could appear across the sets = could carry some bias

General reflections

Mistakes we have made

- Pretraining
  - Should have estimated parameters of data
  - Same person could appear across the sets = could carry some bias
- Finetuning
  - Groundtruth outside of bbox

Conclusion

Succesfully developed and tested models that incorporate temporal smoothing in 2D human pose estimation for bouldering.

- Pretraining and finetuning
- Multiple experiments to find the optimal setting of the models
- The optimal model depends on ones needs

Extras: Mistakes Were Made!

Misimplemented evaluation-function

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	33.3	33.4	32.8	72.5	72.4	73.1	85.8	85.8	86.0
DeciWatch	32.8	33.8	30.9	68.0	68.1	62.7	85.1	84.9	82.8
bi-ConvLSTM - Model S	31.7	30.1	31.6	71.5	68.3	71.3	86.3	82.5	86.2
bi-ConvLSTM - Model C	32.0	32.2	31.8	72.2	72.2	71.4	86.6	86.5	86.5

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	33.1	33.3	32.7	72.1	72.3	72.6	85.6	85.7	85.6
DeciWatch	21.3	26.4	19.4	51.1	58.7	48.1	77.1	81.0	77.0
bi-ConvLSTM - Model S	30.9	31.6	30.6	71.2	71.6	70.7	86.0	86.1	86.2
bi-ConvLSTM - Model C	31.5	32.2	30.8	71.8	72.0	71.9	86.4	86.4	86.5

Extras: Mistakes Were Made!

Misimplemented evaluation-function

1. Decreased accuracies

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	33.3	33.4	32.8	72.5	72.4	73.1	85.8	85.8	86.0
DeciWatch	32.8	33.8	30.9	68.0	68.1	62.7	85.1	84.9	82.8
bi-ConvLSTM - Model S	31.7	30.1	31.6	71.5	68.3	71.3	86.3	82.5	86.2
bi-ConvLSTM - Model C	32.0	32.2	31.8	72.2	72.2	71.4	86.6	86.5	86.5

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	33.1	33.3	32.7	72.1	72.3	72.6	85.6	85.7	85.6
DeciWatch	21.3	26.4	19.4	51.1	58.7	48.1	77.1	81.0	77.0
bi-ConvLSTM - Model S	30.9	31.6	30.6	71.2	71.6	70.7	86.0	86.1	86.2
bi-ConvLSTM - Model C	31.5	32.2	30.8	71.8	72.0	71.9	86.4	86.4	86.5

Extras: Mistakes Were Made!

Misimplemented evaluation-function

1. Decreased accuracies
2. Models do not improve the rough estimates

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	33.3	33.4	32.8	72.5	72.4	73.1	85.8	85.8	86.0
DeciWatch	32.8	33.8	30.9	68.0	68.1	62.7	85.1	84.9	82.8
bi-ConvLSTM - Model S	31.7	30.1	31.6	71.5	68.3	71.3	86.3	82.5	86.2
bi-ConvLSTM - Model C	32.0	32.2	31.8	72.2	72.2	71.4	86.6	86.5	86.5

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	33.1	33.3	32.7	72.1	72.3	72.6	85.6	85.7	85.6
DeciWatch	21.3	26.4	19.4	51.1	58.7	48.1	77.1	81.0	77.0
bi-ConvLSTM - Model S	30.9	31.6	30.6	71.2	71.6	70.7	86.0	86.1	86.2
bi-ConvLSTM - Model C	31.5	32.2	30.8	71.8	72.0	71.9	86.4	86.4	86.5

Extras: Mistakes Were Made!

Misimplemented evaluation-function

1. Decreased accuracies
2. Models do not improve the rough estimates
3. bi-ConvLSTM: Model C is now always better than Model S

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	33.3	33.4	32.8	72.5	72.4	73.1	85.8	85.8	86.0
DeciWatch	32.8	33.8	30.9	68.0	68.1	62.7	85.1	84.9	82.8
bi-ConvLSTM - Model S	31.7	30.1	31.6	71.5	68.3	71.3	86.3	82.5	86.2
bi-ConvLSTM - Model C	32.0	32.2	31.8	72.2	72.2	71.4	86.6	86.5	86.5

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	33.1	33.3	32.7	72.1	72.3	72.6	85.6	85.7	85.6
DeciWatch	21.3	26.4	19.4	51.1	58.7	48.1	77.1	81.0	77.0
bi-ConvLSTM - Model S	30.9	31.6	30.6	71.2	71.6	70.7	86.0	86.1	86.2
bi-ConvLSTM - Model C	31.5	32.2	30.8	71.8	72.0	71.9	86.4	86.4	86.5

Extras: Mistakes Were Made!

Misimplemented evaluation-function

1. Decreased accuracies
2. Models do not improve the rough estimates
3. bi-ConvLSTM: Model C is now always better than Model S
4. 3DConv is generally the best performing model

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	33.3	33.4	32.8	72.5	72.4	73.1	85.8	85.8	86.0
DeciWatch	32.8	33.8	30.9	68.0	68.1	62.7	85.1	84.9	82.8
bi-ConvLSTM - Model S	31.7	30.1	31.6	71.5	68.3	71.3	86.3	82.5	86.2
bi-ConvLSTM - Model C	32.0	32.2	31.8	72.2	72.2	71.4	86.6	86.5	86.5

Accuracy metric	PCK@0.05			PCK@0.1			PCK@0.2		
Experiment	1	2	3	1	2	3	1	2	3
Identity function	19.4	19.4	19.4	66.1	66.1	66.1	85.2	85.2	85.2
3DConv	33.1	33.3	32.7	72.1	72.3	72.6	85.6	85.7	85.6
DeciWatch	21.3	26.4	19.4	51.1	58.7	48.1	77.1	81.0	77.0
bi-ConvLSTM - Model S	30.9	31.6	30.6	71.2	71.6	70.7	86.0	86.1	86.2
bi-ConvLSTM - Model C	31.5	32.2	30.8	71.8	72.0	71.9	86.4	86.4	86.5