Master MVA
Recvis

Vincent Matthys
ENS Paris Saclay

Tools for oral presentation

Vincent Matthys
vincent.matthys@ens-paris-saclay.fr

1 Bibliography

- Linear model to map 3D joints onto 2D joints Martinez, Hossain, Romero, and Little, 2017
- seq2seq model to map 3D joints onto 2D joints Hossain and Little, 2017
- seq2seq model to predict 3D joints positions given action Martinez, Black, and Romero, 2017
- StackedHourglass model Newell, Yang, and Deng, 2016

2 Notations

- N: number of 2D/3D poses
- T: length of the sequence
- $\widehat{Y}_{n,t}$: estimated 3D pose of the n-th joint at time t
- $Y_{n,t}$: ground truth 3D pose of the n-th joint at time t

The MSE over N sequences of T time-steps is given by:

$$\boldsymbol{MSE}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}) = \frac{1}{NT} \sum_{t=1}^{T} \sum_{n=1}^{N} \|\widehat{Y}_{n,t} - Y_{n,t}\|_2^2 \tag{1}$$

Temporal-constraint:

$$\|\Delta \widehat{\boldsymbol{Y}}\|_2^2 = \frac{1}{N(T-1)} \sum_{t=1}^{T-1} \sum_{n=1}^{N} \|\widehat{Y}_{n,t+1} - \widehat{Y}_{n,t}\|_2^2 \tag{2}$$

Overall loss-function:

$$\boldsymbol{\mathcal{L}} = \alpha \boldsymbol{MSE}(\widehat{\boldsymbol{Y}}, \boldsymbol{Y}) + \beta \|\Delta \widehat{\boldsymbol{Y}}\|_2^2 \tag{3}$$

$$\boldsymbol{\mathcal{L}} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\alpha}{T}\sum_{t=1}^{T}\|\widehat{Y}_{n,t} - Y_{n,t}\|_2^2 + \frac{\beta}{T-1}\sum_{t=1}^{T-1}\|\widehat{Y}_{n,t+1} - \widehat{Y}_{n,t}\|_2^2\right) \tag{4}$$

Suggested by Hossain and Little, 2017, due to the difference in radial velocity among joints (thorax and limbs), the temporal-constraint can be expressed in function of considered joints:

$$\|\Delta \widehat{\boldsymbol{Y}}\|_2^2 = \frac{1}{N(T-1)}\sum_{t=1}^{T-1}\sum_{n=1}^{N}\left(\gamma\|\widehat{Y}_{n,t+1}^{trunk} - \widehat{Y}_{n,t}^{trunk}\|_2^2 + \eta\|\widehat{Y}_{n,t+1}^{mid} - \widehat{Y}_{n,t}^{mid}\|_2^2 + \epsilon\|\widehat{Y}_{n,t+1}^{term} - \widehat{Y}_{n,t}^{term}\|_2^2\right) \tag{5}$$

where: (17 joints moving in Human3.6M dataset)

- *trunk* represents head, thorax, hip, l+r hip, spine, neck/nose, l+r shoulder
- *mid* states for r+l knee, r+l elbow
- *term* represents l+r wrist, l+r foot

3 Models

3.1 Linear model

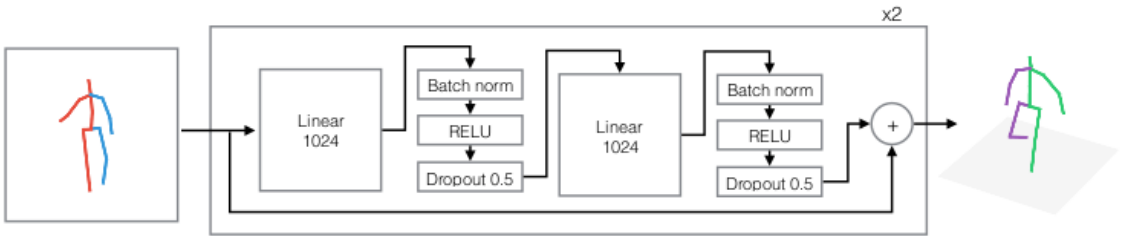

Figure 1: Linear model from (Martinez, Hossain, Romero, & Little, 2017)

Loss-function, basicly MSE:

$$MSE(\widehat{Y}, Y) = \frac{1}{N}\sum_{n=1}^{N}\|\widehat{Y}_n - Y_n\|_2^2 \tag{6}$$

model	direction	discussion	eating	greeting	phoning	photo	posing	purchases
linear	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1
seq2seq	44.2	46.7	52.3	49.3	59.9	59.4	47.5	46.2

model	sitting	sittingdown	smoking	waiting	walkdog	walking	walktogether	Avg
linear	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
seq2seq	59.9	65.5	55.8	50.4	52.3	43.5	45.1	51.9

Table 1: Erros action-wise on Human3.6M dataset from 2D pose estimates from stacked-hourglass network (Newell, Yang, & Deng, 2016) as reported in (Hossain & Little, 2017)

3.2 Sequence-to-sequence model

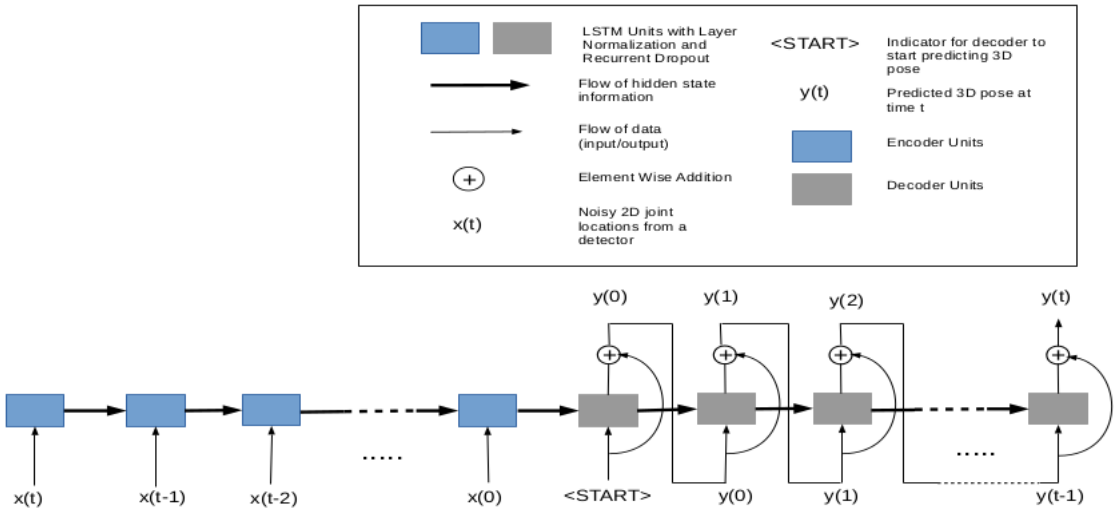

Figure 2: Sequence-to-sequence model from (Hossain & Little, 2017)

Loss-function, MSE over time + temporal constraint:

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\alpha}{T}\sum_{t=1}^{T}\|\widehat{Y}_{n,t} - Y_{n,t}\|_2^2 + \frac{\beta}{T-1}\sum_{t=1}^{T-1}\|\widehat{Y}_{n,t+1} - \widehat{Y}_{n,t}\|_2^2\right) \tag{7}$$

1. Attention mechanism in seq2seq ?
2.

model	direction	discussion	eating	greeting	phoning	photo	posing	purchases
linear	37.7	44.4	40.3	42.1	48.2	54.9	44.4	42.1
seq2seq	35.2	40.8	37.2	37.4	43.2	44.0	38.9	35.6

model	sitting	sittingdown	smoking	waiting	walkdog	walking	walktogether	Avg
linear	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
seq2seq	42.3	44.6	39.7	39.7	40.2	32.8	35.5	39.2

Table 2: Erros action-wise on Human3.6M dataset from 2D ground-truth poses as reported in (Hossain & Little, 2017)

3.3 Results

References

Hossain, M. R. I. & Little, J. J. (2017). Exploiting temporal information for 3d pose estimation. *arXiv preprint arXiv:1711.08585*.

Martinez, J., Black, M. J., & Romero, J. (2017). On human motion prediction using recurrent neural networks. *arXiv preprint arXiv:1705.02445*.

Martinez, J., Hossain, R., Romero, J., & Little, J. J. (2017). A simple yet effective baseline for 3d human pose estimation. *arXiv preprint arXiv:1705.03098*.

Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *European conference on computer vision* (pp. 483–499). Springer.