# Protocol 14
# Model Training

12.04.2019

Two skeleton-based models were used for lameness detection: the Spatial Temporal Graph Convolutional Networks (ST-GCN), and a Resnet. The workflow is demonstrated in Figure 1. The training data contains 501 videos with a frame rate of 20 frames per second, frame size of $680 \times 420$, and length ranging from 4-20 seconds. The poses of the cows were extraxted from the videos, and the coordinates of the pose, or the skeleton joints, are the input for the networks. Lameness detection is considered as a classification problem here, with the locomotion scores (LS) divided into four intervals, each of which is a class:

- Class1: LS 1-2
- Class2: LS 2-3
- Class3: LS 3-4
- Class4: LS 4-5 (containing 5)

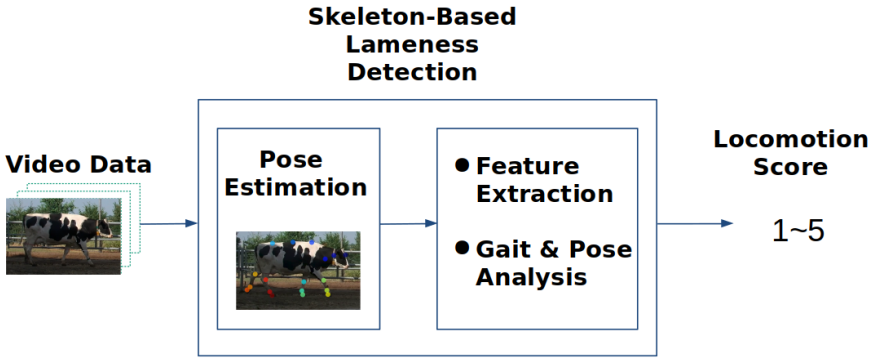


Figure 1: Workflow of skeleton-based lameness detection.

Both models were carried out using Pytorch. The results of both networks show an issue of overfitting, with a test error around 40%.

## 1. Spatial Temporal Graph Convolutional Networks (ST-GCN) [1]

In this type of network, skeleton sequence is be represented by a spatial-temporal graph. In the graph, each node consists of the coordinate of a joint or keypoint. The nodes are connected in such a way that they display the skeleton of human or animal body. The same joint is connected to itself from the neighboring frame, such that the graph is three-dimensional, containing both the spatial and temporal information of body movement. The network was designed for action recognition, and the workflow is shown in Figure 2. The source codes are provided by the authors (https://github.com/yysijie/st-gcn).

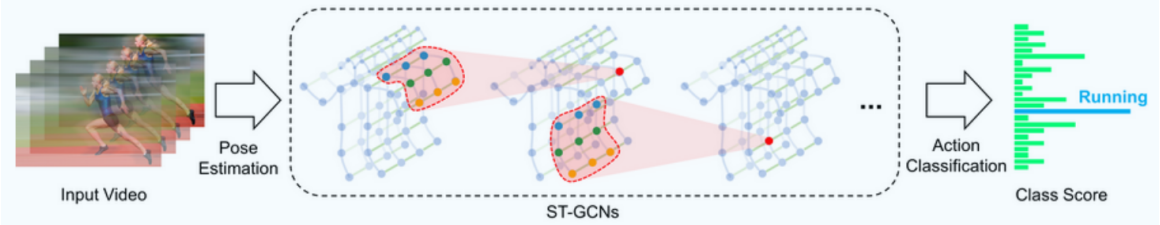


Figure 2: Spatial Temporal Graph Convolutional Network (ST-GCN).

| Batch size | 8 | 16 | 32 | 64 |
|---|---|---|---|---|
| Accuracy | 43.71% | 41.72% | 42.38% | 41.72% |

Table 1: Test accuracy of ST-GCN after 50 epochs using different batch size.

## 2. 2D CNN

The skeleton data were initially transferred to 2D images, which was inspired by [2]. The coordinates of the skeleton joints in different frames form an 2D array, which can be seen as an image, as shown in Figure 3. A pre-trained Resnet was used for lamenes detection.

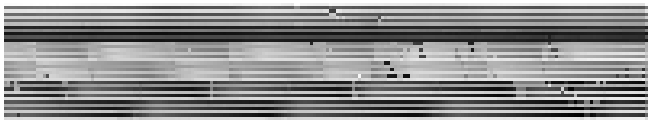

Figure 3: One example of the image of skeleton data.

Figure 4 shows the training curve of the finetuning. While the training loss decreases with the epochs, the test loss does not show much decrease. Hence, some generalization strategies, such as reducing the complexity of the model, should be applied to improve the results. Another issue is if the way of representing skeleton data as an image appropriate, in that some joints are more correlated but are not close to each other in the image.

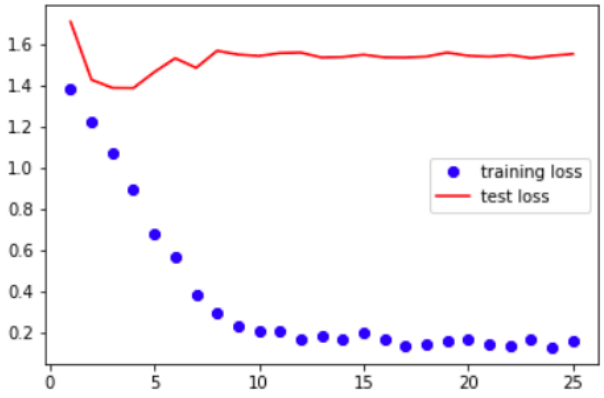


Figure 4: Training curve of Resnet.

## References

[1] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[2] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 597–600, IEEE, 2017.