Protocol 8

There are some neural networks proposed for spatial-temporal feature extraction, mainly used for video analysis, such as action or gait recognition. Even though these networks were used for human action analysis, they can be applied to pose and gait analysis of cows through transfer learning. Three networks are summarized in the following.

1. 3D Convolutional Networks

- **Reference:**

  Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." Proceedings of the IEEE international conference on computer vision. 2015.

- **Input data:** Raw videos

- **Architecture:** 3D CNN: 8 convolution layers, 5 pooling layers, followed by two fully connected layers, and a softmax output layer

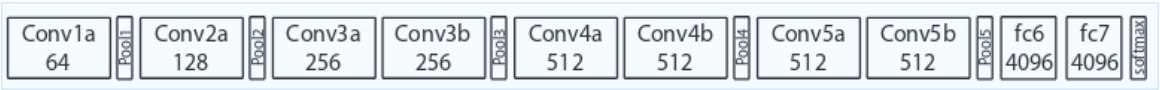

Figure 1: C3D architecture.

- **Traning:**
  - Stochastic gradient descent
  - Data augmentation: flipping, cropping (spatial and temporal jittering)

- **Dataset:** Sports-1M

2. Memory-based Gait Recognition (MGR) Network

- **Reference:**

  Liu, Dan, et al. "Memory-based Gait Recognition." BMVC. 2016.

- **Input data:** 14 * 2D joints for a gait sequence (stacked vectors)

- **Architecture:** LSTMs: 15/8/12 memory blocks with 8/15/8 memory cells

- **Traning:**
  - Leave-one-out cross-validation (one test sample), two-fold cross-validation
  - Data augmentation: removing ten random vectors from gait sequence, quadrupling the original data
  -

- **Dataset:**
  - CASIA A: 240 gait sequences (20 subjects * 4 sequences * 3 views) * 90 frames = 21600 images
  - CASIA B: 13640 gait sequences (124 subjects * 10 sequences * 11 views) * 90 frames = 21600 images

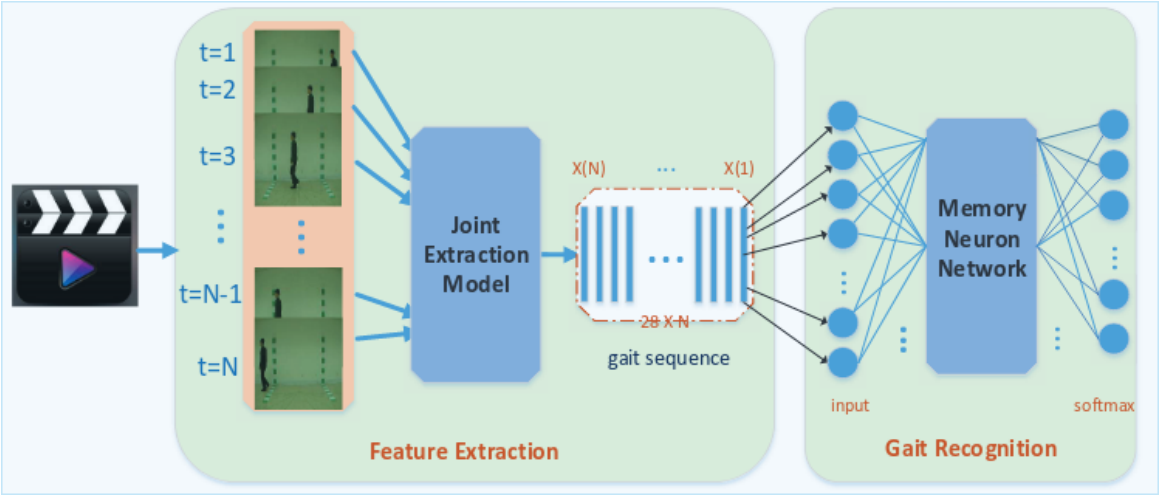

Figure 2: The Memory-based gait recognition framework.

3. Two-Stream Convolutional Network

- **Reference:**

  Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." Advances in neural information processing systems. 2014.

- **Input data:**
  - Single frame for spatial stream
  - Stacked optical flows for temporal stream
- **Architecture:** Two-stream CNN:

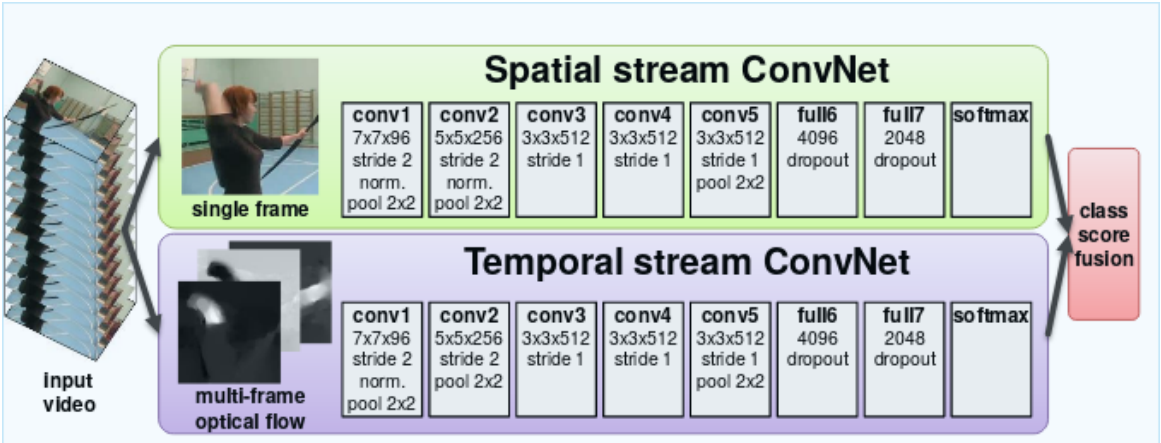

Figure 3: Two-stream architecture for video classification.

- **Traning:**
  - Pre-training: ImageNet ILSVRC-2012
  - mini-batch (256 samples) stochastic gradient descent
  - Data augmentation: cropping, flipping, RGB jittering
  - three-split
- **Dataset:**
  - UCF-101: 13K videos (180 frames/video on average)
  - HMDB-51: 6.8K videos of 51 actions