Multimodal Computational Methods in Political Science

CNNs, RNNs and ViViTs for Video Analysis

Clara Fochler

Ludwig-Maximilians-Universität München

June 30, 2026

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Agenda

Recap: Deep Learning for Motion Estimation

Different CNNs for Video Analysis Models over time

RECAP: RNNs and LSTMs I

RNNs for Video Analysis

Video Vision Transformers (ViViTs)

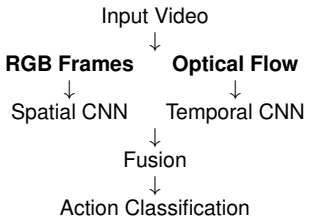
Recap: Deep Learning for Motion Estimation

Coarse-to-fine learning approaches

- SPyNet: pyramid + warping + refinement across scales
- PWC-Net:
  - feature pyramids for both images
  - warping using previous flow estimate
  - cost volume via feature correlation
  - CNN-based refinement per level
  - final context network with dilated convolutions
- Strong correspondence between classical coarse-to-fine methods and deep learning architectures
- Deep models replace explicit energy minimization with learned refinement

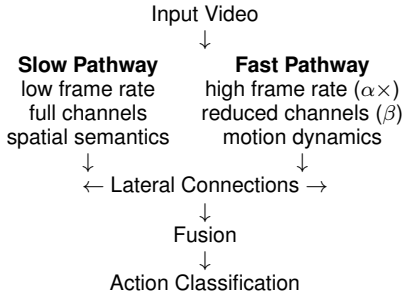
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Two-Stream Network (Simonyan and Zisserman 2014)



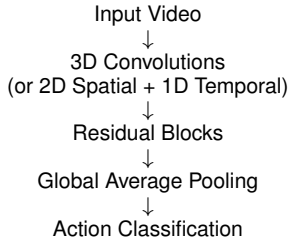
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

SlowFast Networks (Feichtenhofer et al., 2019)



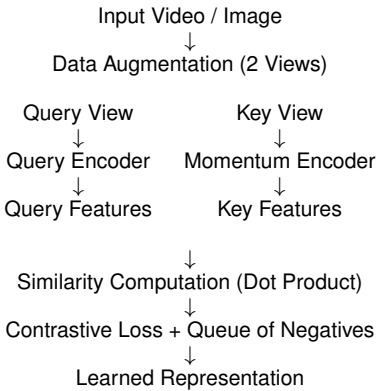
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

3D ResNet (Hara et al. 2017)/ (2+1)D ResNet (Tran et al. 2018)



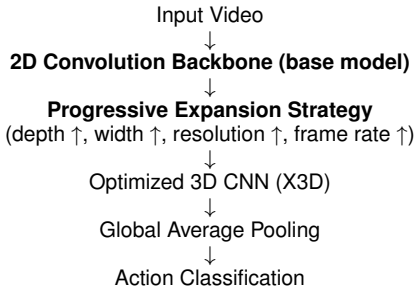
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Self-supervised Models: Momentum Contrast (MoCo) (He et al. 2020)



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

X3D (Feichtenhofer 2020)



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ST-GCN (Yan, Xiong and Lin 2018)

Input Video

↓

Pose Estimation (Skeleton Extraction)

↓

Graph Construction (Joints = Nodes, Bones = Edges)

↓

Spatial-Temporal Graph Convolution (ST-GCN)
(spatial joints (graph convolution) + temporal motion)

↓

Feature Aggregation

↓

Action Recognition

$$\text{new knee} = W_1 \cdot \text{hip} + W_2 \cdot \text{knee} + W_3 \cdot \text{ankle}$$

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

RECAP: RNNs and LSTMs I

RNNs

- Process inputs step-by-step (temporal order matters)
- Maintain a hidden memory of previous inputs

$$
\begin{array}{ccccccccccc}
x_1 & \rightarrow & x_2 & \rightarrow & x_3 & \rightarrow & x_4 & \rightarrow & x_5 \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
h_1 & \rightarrow & h_2 & \rightarrow & h_3 & \rightarrow & h_4 & \rightarrow & h_5
\end{array}
$$

- At each step, the hidden state is updated using the current input and the previous memory

Do remember "The Vanishing Gradient Problem"?

RECAP: RNNs and LSTMs II

LSTMs

- LSTMs are a type of RNN
- Introduce a separate cell state c_t that carries long-term information through time
- Information flow is controlled by **three gates**:
  - **Forget gate:** decides what information to discard from the cell state
  - **Input gate:** decides what new information to store in the cell state
  - **Output gate:** decides what part of the cell state is exposed as hidden state
- The gates are learned functions that allow the model to dynamically control memory
- This structure helps mitigate the vanishing gradient problem by allowing stable information flow through c_t
- LSTMs can model medium-range dependencies (sentences, short paragraphs) but struggle with very long documents

What are shortcomings of CNNs that RNNs might compensate in the context of Video Analysis?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

RNNs for Video Analysis

We use RNNs (mostly LSTMs) in combination with CNNs to compensate for CNNs short-coming in capturing temporal/sequence of frames

LRCN (Long-term Recurrent ConvNet) (Donahue et al. 2016)
Input Video → CNN (frame features) → LSTM → Class label

ConvLSTM (Shi et al. 2015)
Spatio-temporal input (e.g. radar / frame sequence)
→ ConvLSTM (joint spatial + temporal memory)
→ Prediction map

Unsupervised Video Representation (Srivastava et al. 2015)
Unlabeled video clips → CNN + LSTM encoder (feature extraction)
→ Soft-attention over time → Latent video embedding

Video Captioning (Encoder–Decoder) (Yao et al. 2015)
Input Video → CNN/LSTM encoder (features)
→ LSTM decoder (language generation)
→ Generated caption

- LSTMs still do not hold information for long sequences
- Training and inference are slow - why?

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Video Tokenization

Uniform Frame Sampling

- **Approach:** Uniformly sample n_t frames from a video clip $V \in \mathbb{R}^{T \times H \times W \times C}$.
- **Extraction:** Each 2D frame of resolution $H \times W$ is split into patches of size $P \times P$ and flattened into vectors $x_p \in \mathbb{R}^{N_{frame} \times (P^2 \cdot C)}$
- **Tokens:** Total tokens = $n_t \times \frac{H}{P} \times \frac{W}{P}$. Temporal fusion occurs later within the transformer encoder

Tubelet Embedding (3D Convolution)

- **Approach:** Extract non-overlapping spatiotemporal "tubes" of size $t \times p \times p$ directly from the volume V
- **Extraction:** Tubes capture both spatial patches and temporal context, then flattened to build video tokens
- **Tokens:** Total tokens = $\frac{T}{t} \times \frac{H}{p} \times \frac{W}{p}$. Spatiotemporal information is fused directly during tokenization

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ViViT: Spatio-temporal attention (Arnab et al., 2021)

Model 1: Joint Space-Time Attention

Input Video

↓

Patch Embedding + Positional Encoding (spatio-temporal)

↓

Single Transformer Encoder $\times L$

(global self-attention over all tokens: space + time)

↓

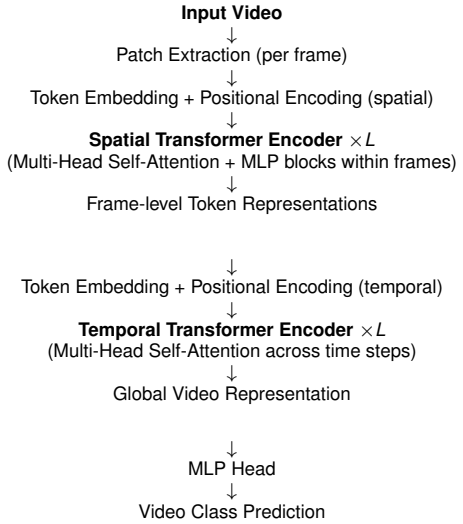
Global Video Representation

↓

MLP Head $\rightarrow$ Class

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ViViT: Factorised Encoder (Arnab et al., 2021)



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ViViT: Factorised Self-Attention

Input Video

↓

Patch Embedding

↓

Spatial Self-Attention

(attention within each frame, same time index)

↓

Intermediate token representations

↓

Temporal Self-Attention

(attention across frames, same spatial index)

↓

Global Video Representation

↓

MLP Head $\rightarrow$ Class

No CLS token used due to ambiguity in token aggregation

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

ViViT: Factorised Dot-Product Attention

Input Video
↓
Patch Embedding
↓
Multi-Head Self-Attention Split
↓
Half of heads: Spatial attention (within frames)
Half of heads: Temporal attention (across frames)
↓
Fused token representations
↓
Global Video Representation
↓
MLP Head $\rightarrow$ Class

Efficient factorisation of attention heads

CNNs for Political Science: Legislative Speech Analysis

Rittmann, O., Ringwald, T., Nyhuis, D. (2025). Public Opinion and Emphatic Legislative Speech: Evidence from an Automated Video Analysis. British Journal of Political Science, 55, e165

RQ: When do politicians deliver more emphatic speeches?

Data

- recordings of key debates in the House of Representatives, camera fully captures the speaker, 2,341 speeches by 543 legislator, $\approx$16 min each $\rightarrow$ 6.4M frames
- Stratified training sample: 245 speeches (184 train / 61 test), varying emphasis levels

Method

- Pseudo-3D ResNet (similar to (2+1)D ResNet + pretrained on Kinetics) fine-tuned on speech emphasis
- Multimodal: audio (but separate soundnet-like subnetwork) + video streams
- dropout against overfitting

Finding

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

REFERENCES I

- Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C. (2021). ViViT: A Video Vision Transformer (arXiv:2103.15691). arXiv. https://doi.org/10.48550/arXiv.2103.15691
- Feichtenhofer, C., Fan, H., Malik, J., He, K. (2019). SlowFast Networks for Video Recognition (arXiv:1812.03982). arXiv. https://doi.org/10.48550/arXiv.1812.03982
- Feichtenhofer, C. (2020). X3D: Expanding Architectures for Efficient Video Recognition (arXiv:2004.04730). arXiv. https://doi.org/10.48550/arXiv.2004.04730
- Hara, K., Kataoka, H., Satoh, Y. (2017). Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition (arXiv:1708.07632). arXiv. https://doi.org/10.48550/arXiv.1708.07632
- He, K., Fan, H., Wu, Y., Xie, S., Girshick, R. (2020). Momentum Contrast for Unsupervised Visual Representation Learning (arXiv:1911.05722). arXiv. https://doi.org/10.48550/arXiv.1911.05722

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

REFERENCES II

- Hugging Face. (2025). Community Computer Vision Course: Unit 7 – Video and Video Processing. https://huggingface.co/learn/computer-vision-course/unit7/video-processing/introduction-to-video
- Nyhuis, D., Ringwald, T., Rittmann, O., Gschwend, T., Stiefelhagen, R. (2021). Automated video analysis for social science research. In U. Engel, A. Quan-Haase, S. X. Liu, L. Lyberg (Eds.), Handbook of Computational Social Science, Volume 2: Data science, statistical modelling, and machine learning methods (pp. 386–398). Routledge. https://doi.org/10.4324/9781003025245-26
- Rittmann, O., Ringwald, T., Nyhuis, D. (2025). Public Opinion and Emphatic Legislative Speech: Evidence from an Automated Video Analysis. British Journal of Political Science, 55, e165
- Simonyan, K., Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos (arXiv:1406.2199). arXiv. https://doi.org/10.48550/arXiv.1406.2199

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

REFERENCES III

- Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M. (2018). A Closer Look at Spatiotemporal Convolutions for Action Recognition (arXiv:1711.11248). arXiv. https://doi.org/10.48550/arXiv.1711.11248
- Yan, S., Xiong, Y., Lin, D. (2018). Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition (arXiv:1801.07455). arXiv. https://doi.org/10.48550/arXiv.1801.07455

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Thanks for your attention!