

Computer Vision - 2025

Week #11. Video Data Processing

Lectures by Alexei Kornaev ^{1,2,3}

Practical sessions by Kirill Yakovlev ²

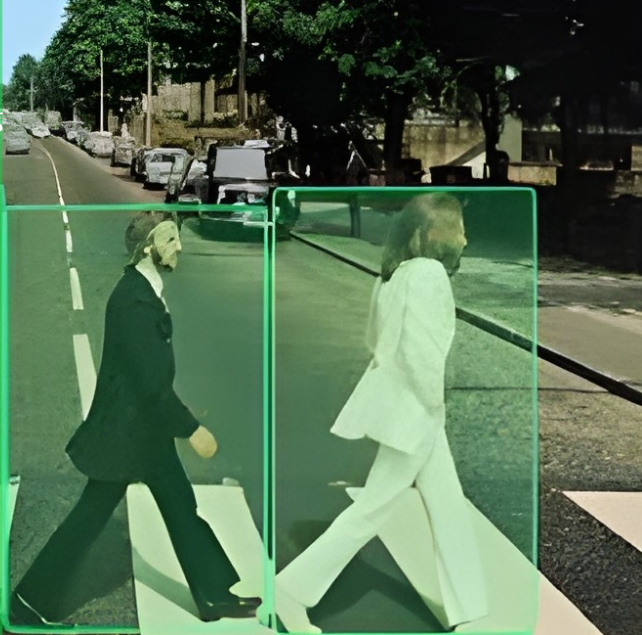
¹AI Institute, Innopolis University (IU), Innopolis

²Robotics & CV Master's Program, IU, Innopolis

³Dept. of M^2R , Orel State University, Orel

⁴RC for AI, National RC for Oncology, Moscow

March 31, 2025



Agenda

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

- ① Outcomes
- ② Theoretical Background and Historical Context
- ③ Video Processing with ML
- ④ Conclusion

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

Section 1. Outcomes

Outcomes

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

This week's lecture and practical session on Video Data Processing aim to provide an understanding of motion analysis, temporal modeling, and deep learning-based video processing. By the end of this week, students will be able to:

- 1 Understand the principles of optical flow, video stabilization, action recognition, and video summarization.
- 2 Describe different temporal modeling approaches, including 2D CNNs + RNNs (LSTMs), 3D CNNs, and Video Transformers (ViViT).
- 3 Implement a video-based classification/approximation model using 2D CNNs, 2D CNN + LSTM, 3D CNNs, and ViViT.

Key Takeaway: Video data processing is essential for applications in robotics, surveillance, autonomous driving, and human activity recognition, with deep learning playing a crucial role in modern advancements.

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

Section 2. Theoretical Background and Historical Context

Optical Flow Fundamentals

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

Definition & Importance

- **Optical Flow:** The apparent motion of brightness patterns in a sequence of images.
- Provides information about object movement and scene structure.
- Critical for tasks such as motion estimation, tracking, and video stabilization.

Mathematical Formulation

$$\frac{\partial I}{\partial x} u + \frac{\partial I}{\partial y} v + \frac{\partial I}{\partial t} = 0, \quad (1)$$

where u and v denote the horizontal and vertical flow components.

Historical Context

- Early work by [Horn & Schunck](#) [Horn and Schunck \[1981\]](#) laid the foundation.
- Nowadays the optical flow methods are implemented with DL models, e.g. RAFT.

Video Stabilization Techniques

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

Objective

- Remove unwanted camera motion (e.g., shake) to produce smooth videos.
- Enhance the visual quality and reliability for further processing.

Methodologies

- Feature tracking and estimation of global motion.
- Optimization techniques using robust norms (e.g., L1 minimization).

Notable Contribution

- Auto-directed video stabilization with robust L1 optimal camera paths ?.

Action Recognition & Video Summarization

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

Action Recognition

- Transition from handcrafted features to deep learning approaches.
- Two-stream CNNs capture spatial and temporal information ?.
- 3D CNNs extend the concept to spatio-temporal feature extraction ?.

Video Summarization

- Identifying keyframes and key events from long video sequences.
- Balancing unsupervised vs. supervised approaches ?.

Method	Temporal Modeling	Reference
Two-Stream CNN	Frame-wise + Temporal Fusion	?
3D CNN	Direct Spatio-temporal Extraction	?
Summarization	Keyframe/Event Extraction	?

Table: Comparative Overview of Methods

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

Section 3. Video Processing with ML

Recap: 2D Convolution Applied to an Image

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

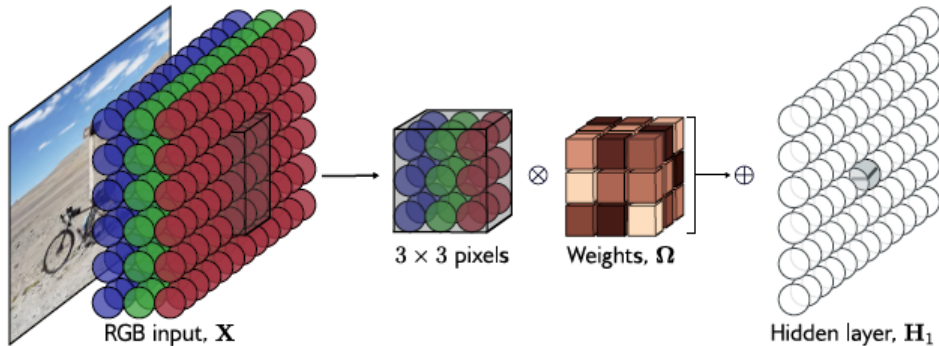


Figure: The image is treated as a 2D input with three channels corresponding to the red, green, and blue components. With a 3×3 kernel, each pre-activation in the first hidden layer is computed by point-wise multiplying the $3 \times 3 \times 3$ kernel weights with the 3×3 RGB image patch centered at the same position, summing, and adding the bias. To calculate all the pre-activations in the hidden layer, we "slide" the kernel over the image in both horizontal and vertical directions. The output is a 2D layer of hidden units. To create multiple output channels, we would repeat this process with multiple kernels, resulting in a 3D tensor of hidden units at hidden layer H_1 [Prince, 2023].

Paper reading

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

ViViT: A Video Vision Transformer

We present pure-transformer based models for video classification, drawing upon the recent success of such models in image classification. Our model extracts spatio-temporal tokens from the input video, which are then encoded by a series of transformer layers. In order to handle the long sequences of tokens encountered in video, we propose several, efficient variants of our model which factorise the spatial- and temporal-dimensions of the input. Although transformer-based models are known to only be effective when large training datasets are available, we show how we can effectively regularise the model during training and leverage pretrained image models to be able to train on comparatively small datasets. We conduct thorough ablation studies, and achieve state-of-the-art results on multiple video classification benchmarks including Kinetics 400 and 600, Epic Kitchens, Something-Something v2 and Moments in Time, outperforming prior methods based on deep 3D convolutional networks. To facilitate further research, we will release code and models [Arnab et al. \[2021\]](#).

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

Section 4. Conclusion

Conclusion

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

In this lecture, we explored the fundamental concepts and modern advancements in video data processing. We covered:

- Optical Flow: Classical methods (Lucas-Kanade, Horn-Schunck) and deep learning-based models (FlowNet, PWC-Net, RAFT).
- Video Stabilization & Action Recognition: Key techniques for video preprocessing and activity classification.
- Temporal Modeling: Comparing RNNs, LSTMs, 3D CNNs, and Video Transformers.
- Hands-on Applications: Training models for classification.

Key Takeaway: Modern video processing leverages deep learning to improve motion estimation, activity recognition, and temporal modeling. Transformers are emerging as a powerful tool for video understanding.

Bibliography

CV-2025

A.Kornaev,
K.Yakovlev

Outcomes

Theoretical
Background
and Historical
Context

Video
Processing
with ML

Conclusion

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021. URL <https://arxiv.org/abs/2103.15691>.

Berthold KP Horn and Brian G Schunck. Determining optical flow. **Artificial Intelligence**, 17(1-3):185–203, 1981.

Simon J.D. Prince. **Understanding Deep Learning**. The MIT Press, 2023. URL <http://udlbook.com>.