Hybrid CNN–RNN Framework for Automatic Video Summarization

Course: **ANN & DL**
Course code: **AI253IA**

Presented by

L Moryakantha	1RV24AI406
Mahantesh	1RV24AI407

Problem Statement

The exponential growth of video content across platforms like surveillance, education, and entertainment has created a pressing need for efficient summarization methods. Existing systems either rely solely on CNNs—capturing spatial features but ignoring temporal dependencies—or on RNNs—focusing on sequence modeling but lacking visual understanding. As a result, current models produce summaries that are often redundant, contextually weak, or miss key events.

This project addresses this gap by developing a hybrid CNN–RNN framework capable of learning both spatial and temporal features, enabling the generation of concise, meaningful, and context-aware video summaries.

Introduction

- In the era of digital transformation, video has become one of the most dominant forms of data across various domains such as surveillance, education, entertainment, and social media. The exponential increase in video content has created a significant challenge in terms of efficient storage, retrieval, and analysis. Manually reviewing or annotating long videos is time-consuming and often impractical, which highlights the need for automated video summarization systems capable of generating concise yet meaningful summaries.

- Traditional approaches to video summarization often rely on either Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) independently. CNNs are effective at extracting spatial features—understanding objects and scenes within individual frames—while RNNs excel at capturing temporal dependencies—the sequential flow of events over time. However, when used separately, these models fail to fully capture both spatial and temporal contexts simultaneously, leading to incomplete or redundant summaries.

- This project proposes a Hybrid CNN–RNN Video Summarization Framework that integrates the strengths of both architectures. The CNN component extracts visual features from each frame, and the RNN component models temporal relationships across frames to predict frame-level importance. The system aims to produce summaries that are context-aware, visually coherent, and time-efficient, suitable for applications in surveillance footage analysis, lecture video summarization, and media content compression.

- By combining deep learning techniques with efficient preprocessing and evaluation strategies, the proposed system addresses current limitations and contributes toward the development of intelligent, automated, and adaptive video summarization solutions.

Literature Review

Referen ces	Title	Authors & Date	Information from Reference
[1]	Video Summarization Using Deep Neural Networks: A Survey	Evlampios Apostolidis, Eleni Adamantidou, Anastasios Mezaris, Ioannis Kompatsiaris (2021)	Comprehensive survey of video summarization techniques, covering CNN, RNN, GAN, RL, and attention-based methods; provides taxonomy and challenges.
[2]	Video Summarization with Attention-Based Encoder-Decoder Networks	Zhong Ji, Kun Xu, Yanwei Pang, Jungong Han (2018)	Proposes an encoder-decoder model combining CNN frame encoding with RNN-based sequence modeling, enhanced by attention to capture keyframes.
[3]	Unsupervised Video Summarization with Adversarial LSTM Networks	Behrooz Mahasseni, Michael Lam, Sinisa Todorovic (2017)	Introduces an adversarial LSTM (RNN + GAN) for unsupervised summarization, learning compact summaries without manual annotations
[4]	Video Summarization with Long Short-Term Memory	Ke Zhang, Wei-Lun Chao, Fei Sha, Kristen Grauman (2016)	Classic supervised model combining CNN features with bi-directional LSTM and diversity optimization (DPP)

Literature Review

Referen ces	Title	Authors & Date	Information from Reference
[5]	A Video Summarization Model Based on Deep Reinforcement Learning with Long-Term Dependency	Xu Wang, Xiaolin Hu, Peng Zhang (2022)	Employs deep reinforcement learning combined with CNN-LSTM to optimize long-term dependency and summary quality
[6]	Sequence to Sequence – Video to Text	Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney (2015)	Pioneer work applying CNN-LSTM for video captioning, converting frame sequences into descriptive sentences
[7]	Summarizing Videos with Attention	Jiří Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso (2019)	Introduces self-attention mechanisms (Transformer-style) for identifying salient frames in videos without recurrent layers
[8]	Progressive Video Summarization via Multimodal Self-Supervised Learning	Yuxuan Li, Yang Liu, Jingen Liu (2023)	Latest multimodal model that fuses visual + audio features with CNN-LSTM and self-supervised learning to generate summaries progressively

Methodology / Proposed System

1. **Environment Setup and Data Collection**
   - The development environment is initialized using TensorFlow/Keras or PyTorch with support libraries like OpenCV and NumPy.
   - Each video is preprocessed — frames are extracted at regular intervals, resized, normalized, and stored for feature extraction.

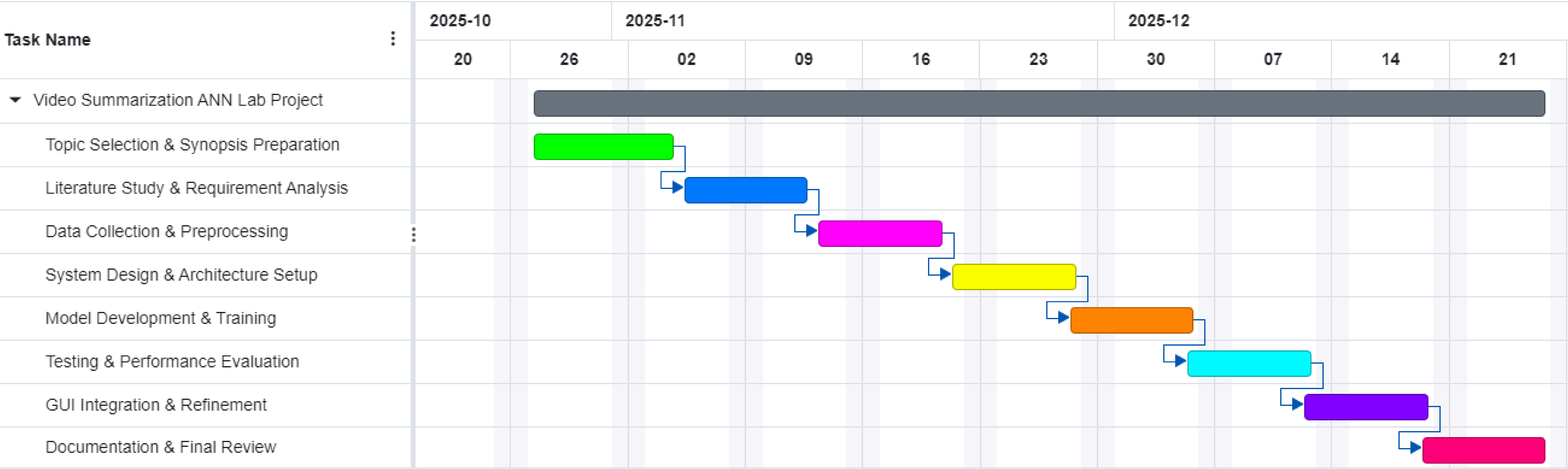
2. **Model Design and Training**
   - A Hybrid CNN–RNN architecture is implemented:
     - CNN Backbone (e.g., VGG16, ResNet50) for spatial feature extraction from individual frames.
     - RNN Module (LSTM/GRU) for temporal dependency modeling across frames.
   - Training is performed using a supervised or weakly-supervised loss function based on ground truth importance scores.
   - Optimization is done via Adam optimizer with learning rate scheduling and early stopping to prevent overfitting.

3. **Evaluation and Benchmarking**
   - The trained model is evaluated on unseen videos from the dataset.
   - Performance metrics include Precision, Recall, F-score, and Summarization Ratio compared against ground-truth summaries.
   - Visualization modules plot learning curves (loss vs. epochs) and show predicted vs. actual summary segments.

4. **Post-processing and Summary Generation**
   - The frame importance scores are thresholded to select keyframes or key segments.
   - Selected frames are stitched together using OpenCV to form the final video summary.

5.**Deployment / Demonstration**
   - The model is deployed in a streamlined video summarization interface for real-time or batch processing.
   - Users can upload videos and automatically receive short summaries highlighting key events.

Timeline – Gantt Chart



Expected Outcomes

- A functional prototype will be developed using a Hybrid CNN–RNN architecture for automated video summarization. The system will demonstrate the ability to learn both spatial and temporal patterns from video data, effectively identifying keyframes and generating concise summaries that preserve contextual meaning. The prototype will include real-time visualization of frame selection, importance scoring, and summary generation, providing an interpretable understanding of how the model prioritizes visual information.

- Performance will be evaluated using quantitative metrics such as Precision, Recall, F-score, and Summarization Ratio. Comparative benchmarks will be drawn between single-model baselines (CNN-only or RNN-only) and the proposed hybrid framework, emphasizing improvements in temporal coherence, content coverage, and reduction of redundancy.

- Future extensions include integrating Transformer-based attention mechanisms for enhanced contextual understanding, deploying the model for real-time summarization in surveillance or lecture environments, and expanding to multi-modal summarization (audio–visual–text). Further enhancements may involve cloud-based deployment, user-interactive summarization dashboards, and adaptive learning for personalized content summarization.

References

- [1] **[2]** [2101.06072] **Video Summarization Using Deep Neural Networks:** A Survey
- https://arxiv.org/abs/2101.06072
- [3] [4] **Video Summarization Techniques: A Comprehensive Review**
- https://arxiv.org/html/2410.04449v1
- [5] [6] arxiv.org
- https://arxiv.org/pdf/1708.09545
- [7] [8] **Unsupervised Video Summarization With Adversarial LSTM Networks**
- https://openaccess.thecvf.com/content_cvpr_2017/papers/Mahasseni_Unsupervised_Video_Summarization_CVPR_2017_paper.pdf
- [9] cs.utexas.edu
- https://www.cs.utexas.edu/~grauman/papers/zhang-eccv2016-lstm-summ.pdf
- [10] [11] **A Video Summarization Model Based on Deep Reinforcement Learning with Long-Term Dependency**
- https://www.mdpi.com/1424-8220/22/19/7689
- [12] arxiv.org
- https://arxiv.org/pdf/1812.01969
- [13] [1505.00487] **Sequence to Sequence -- Video to Text**
- https://arxiv.org/abs/1505.00487
- [14] **Progressive Video Summarization via Multimodal Self-Supervised Learning**
- https://openaccess.thecvf.com/content/WACV2023/papers/Li_Progressive_Video_Summarization_via_Multimodal_Self-Supervised_Learning_WACV_2023_paper.pdf
- [15] **Long-Term Recurrent Convolutional Networks for Visual Recognition and Description**
- https://openaccess.thecvf.com/content_cvpr_2015/papers/Donahue_Long-Term_Recurrent_Convolutional_2015_CVPR_paper.pdf
- [16] **Multi-Annotation Attention Model for Video Summarization**
- https://openaccess.thecvf.com/content/CVPR2023W/LSHVU/papers/Terbouche_Multi-Annotation_Attention_Model_for_Video_Summarization_CVPRW_2023_paper.pdf
- [17] [1801.00054] **Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward**
- https://arxiv.org/abs/1801.00054
- [18] CFSum: **A Transformer-Based Multi-Modal Video Summarization Framework With Coarse-Fine Fusion**
- [19] **AI-driven video summarization for optimizing content retrieval and management through deep learning techniques** | Scientific Reports
- https://www.nature.com/articles/s41598-025-87824-9?error=cookies_not_supported&code=94a61406-2e5a-4d14-a779-947d0d00a09e
- [20] **VideoSAGE: Video Summarization with Graph Representation Learning**
- https://openaccess.thecvf.com/content/CVPR2024W/SG2RL/papers/Chaves_VideoSAGE_Video_Summarization_with_Graph_Representation_Learning_CVPRW_2024_paper.pdf
- [21] **FullTransNet: Full Transformer with Local-Global Attention for Video Summarization**
- https://arxiv.org/pdf/2501.00882
- [22] **Weakly Supervised Deep Reinforcement Learning for Video Summarization With Semantically Meaningful Reward**
- https://openaccess.thecvf.com/content/WACV2021/papers/Li_Weakly_Supervised_Deep_Reinforcement_Learning_for_Video_Summarization_With_Semantically_WACV_2021_paper.pdf
- [23] **Weakly-supervised Video Summarization using Variational Encoder-Decoder and Web Prior**
- https://openaccess.thecvf.com/content_ECCV_2018/papers/Sijia_Cai_Weakly-supervised_Video_Summarization_ECCV_2018_paper.pdf
- [24] [25] **CSTA: CNN-based Spatiotemporal Attention for Video Summarization**
- https://openaccess.thecvf.com/content/CVPR2024/papers/Son_CSTA_CNN-based_Spatiotemporal_Attention_for_Video_Summarization_CVPR_2024_paper.pdf
- [26] **SummDiff: Generative Modeling of Video Summarization with Diffusion**
- https://arxiv.org/html/2510.08458v1
- [27] **Unsupervised Video Summarization via Iterative Training and Simplified GAN**
- https://arxiv.org/html/2311.03745v2
- [28] [2306.01395] **Masked Autoencoder for Unsupervised Video Summarization**
- https://arxiv.org/abs/2306.01395
- [29] iti.gr
- https://www.iti.gr/~bmezaris/publications/csvt20_preprint.pdf
- [30] **Human-Inspired Summarization: Cluster Scene Videos into Diverse Frames**
- https://arxiv.org/html/2311.17940v2

THANK YOU