UNIVERSITY OF ALBERTA

Automatically identifying action sequences in movies

20th January, 2025

Presented by : Abhishek Sharma & Alton Dsouza

1. Traditional Approach - Analysis, Transcription & Subtitle Based Approaches

Action movies segmentation and summarization based on tempo analysis

Learning Realistic Human Actions from Movies

Mining visual actions from movies

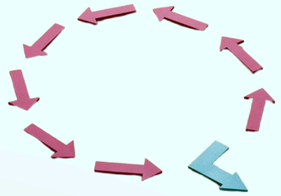
Continuous Action Recognition Based on Sequence Alignment

Rough PipeLine Overview

Baseline

3. Shot Extraction

4. Shot Classification & Annotation



2. Subtitle & Transcription Alignment

5. Tempo Analysis

1. Dataset Collection : Subtitle & Transcription extraction.

7. Performance Evaluation against Baseline Models

6. Data Collation

2. DeepLearning Based Approach - Analysis, Multi modal based approaches using Deep Neural Networks e.g. LSTM, CNN.

A. Deep Analysis of CNN-based Spatio-temporal Representations for Action Recognition

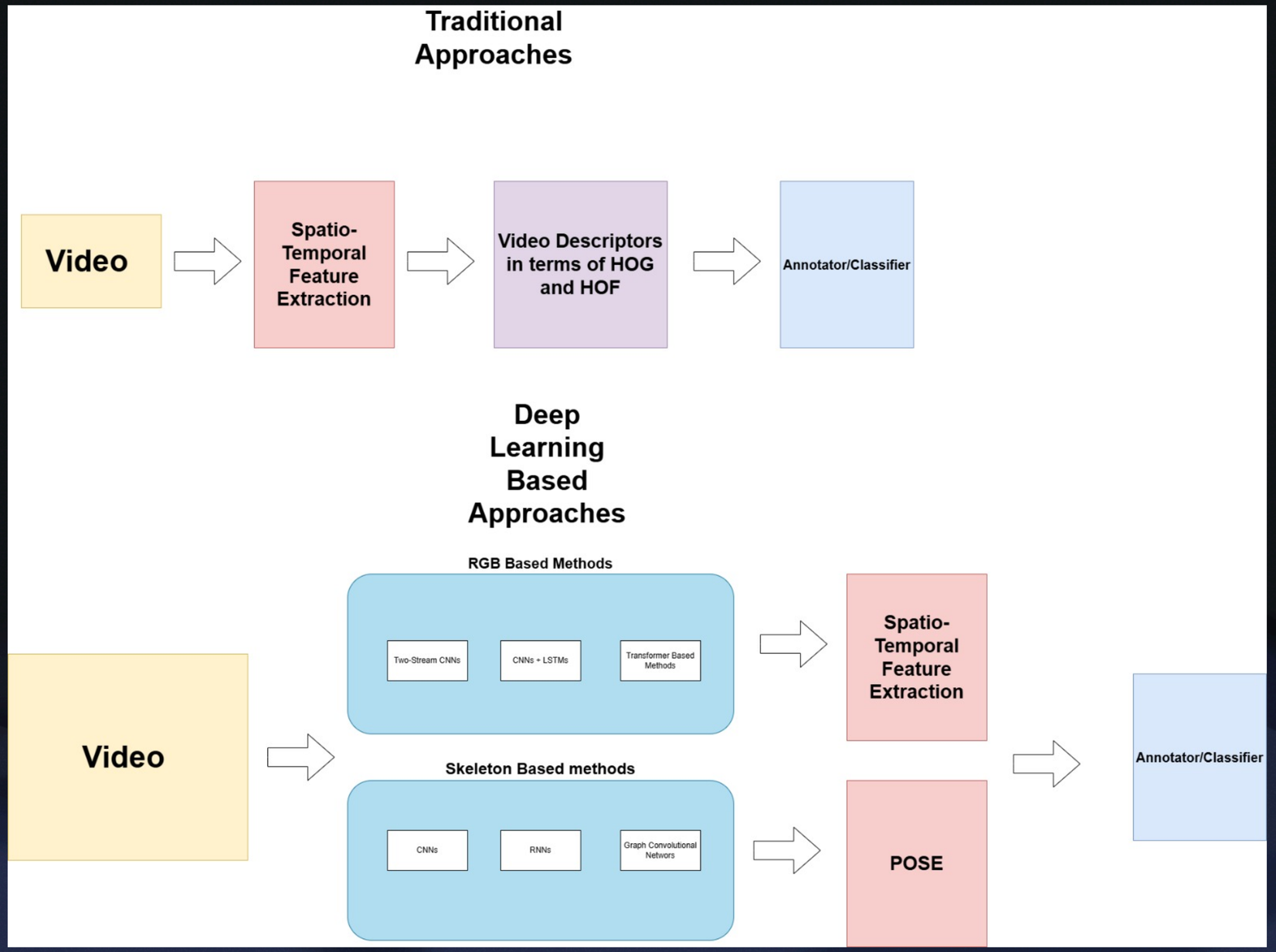
B. A review of action recognition based on Convolutional Neural Network

C. A Review of Machine Learning and Deep Learning for Object Detection, Semantic Semantic Segmentation, and Human Action Recognition in Machine and Robotic Vision

D. Review of Human Action Recognition Based On Deep Learning

E. The Journey of Action Recognition (pre-print)

1. Traditional Approach - Analysis, Transcription & Subtitle Based Approaches



VLM & LLM Based Approach - Transformer Based Approach, e.g Blip, OFA-large, VQA, NLP, MultiModal, CNN etc.

A. Are Visual Language Models Effective iin Action Recognition? A Comparative Study.

B. Large Language Models are Visual Reasoning Coordinators Can VLMs be used on videos for action recognition?

C. ActionCLIP: Adpating Language Image Pretrained Models for video Action Recognition

D. ViLP: Knowledge Exploration using Vision, Language, and Pose Embeddings for Video Action Recognition

E. Language Model Guided Interpretable Video Action Reasoning

F. Movie Clip visual Scene recognition on movies

G. Actom Sequence Models for Efficient Action Detection

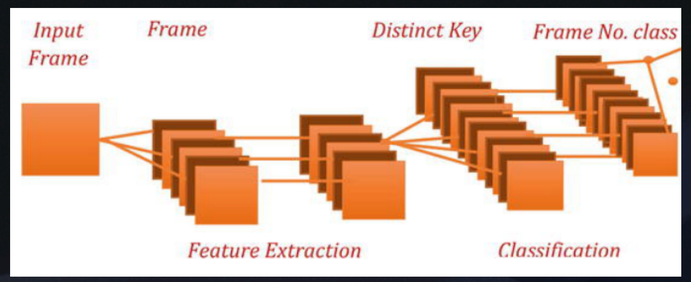
H. Overview of temporal action detection based on deep learning

I. STEP: Spatio-Temporal Progressive Learning for Video Action Detection

Learning Realistic Human Actions from Movies

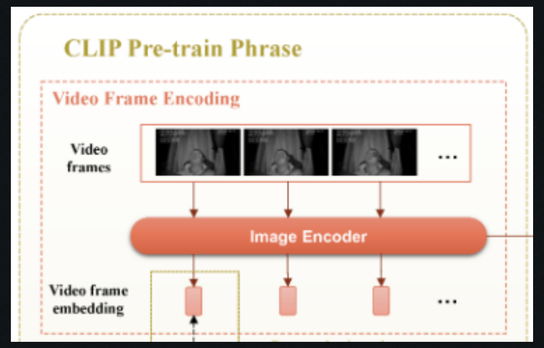
UNIVERSITY OF ALBERTA

Input Video or Image data from datasets



Frame Extraction

Feature Extraction



Visual Encoders for spatial and temporal feature extraction

Use textual encoders (e.g., CLIP's text encoder, FLAN-T5) for processing action labels, captions, or queries.

Optional modalities (e.g., pose heatmaps, optical flow) are extracted for additional context.

Multi Modal Alignment : Align Visual, textual and Optional modality in a joint space

Learning : Joint embedding space alignment to connect video features with textual description

Temporal modeling for capturing sequential information

Cross Modal learning

Zero Shot or Few shot classification

Multi Label classification, action segmentation, or genre prediction

Visual Questioning answering or analysis

UNIVERSITY OF ALBERTA

Thank you