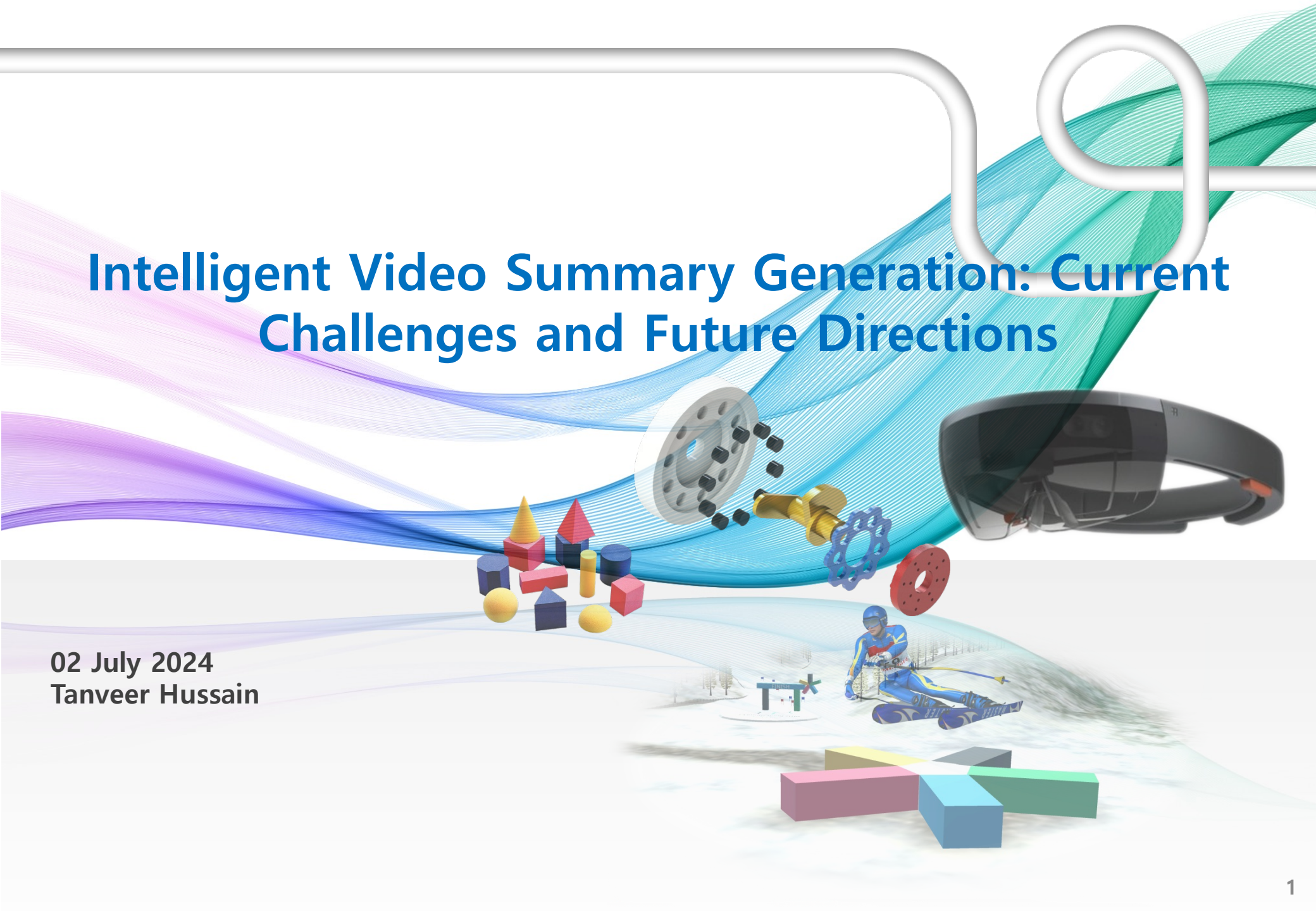
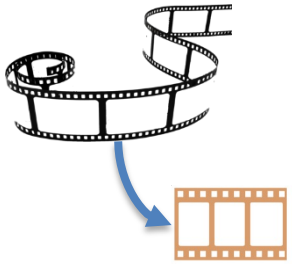
Intelligent Video Summary Generation: Current Challenges and Future Directions



02 July 2024

Tanveer Hussain

INDEX



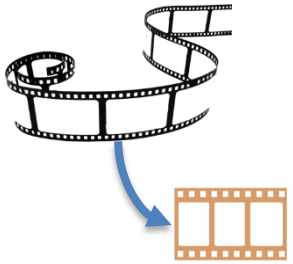
Main Contents

INDEX

01

I. Introduction

Video Summarization Overview
Motivation
Applications



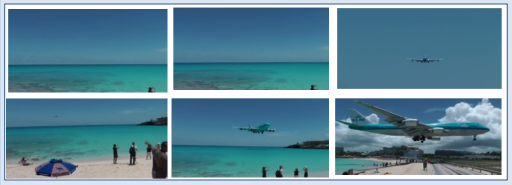
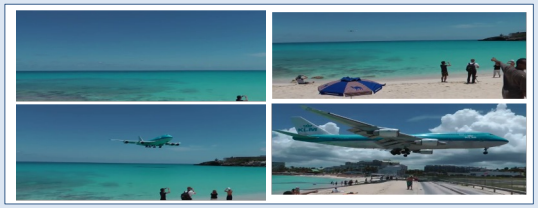
Video Summarization Overview

- "Give a brief statement of the main points of (something)" [Oxford dictionary]
- A video is summarized by extracting the most important contents (keyframes) while eliminating redundant frames.

Pipeline of video summarization

Input	Processing	Output	
Video	Video Summarization	Video Skim	Keyframes
		Short duration video	Representative frames

Main stages of video summarization

Preprocessing	Redundancy Removal	Keyframes Selection
		

Characteristics of an ideal video summary

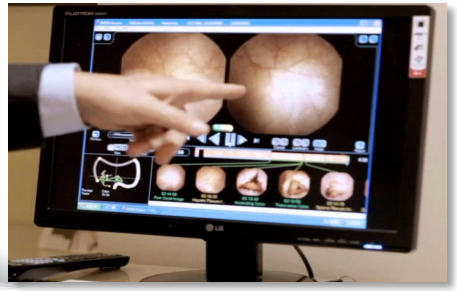
Non-redundant	Semantically significant	Concise	Continuity	Diverse

Motivation for Video Summarization

- Exponential increase in the amount of video data generated regularly
- Huge volume, diversity, and redundancy limit their usefulness in practical applications

Surveillance
Medical Video Repositories
Videos on Social Media
Tourist Videos

Major Sources for Video Generation


Video Data Usage Statistics

1 →	**Fast Networks** Fast data networks allow efficient transfer of video data	82% web traffic is for video contents for 2019
2 →	**Extensive Video Sharing and Social Networking Websites** Rising tendency to share videos on the social web	More than 4 million hours uploaded to YouTube daily (2016)
3 →	**Cheap Storage Devices** Advanced and inexpensive devices allow us to store huge volumes of data	Facebook process 5PB data daily (2016)
4 →	**Mobile Computing** Prevailing use of mobile devices to perform day-to-day tasks	92% mobile viewers share videos

→ Video summarization can be used to remove redundancy from videos and generate their condensed representations for efficient processing

Video Summarization Applications

Entertainment

- Movie trailers

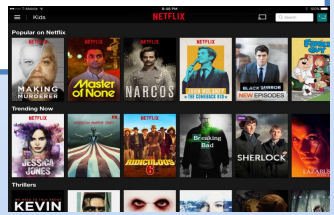
→ **Compact representation of original video to grab viewer's interest**

- Video libraries

→ **Summarized video acts as index for original video in libraries**

- Movie recommendation

→ **Summarization allows efficient video genre classification**

Sports

- Interesting events based summarization

→ **Construct short video skims of interesting events like attempts at goal in a soccer game**

- Effectively shorten lengthy events for quick viewing

→ **Enable users to quickly view the most interesting events in lengthy sports videos**



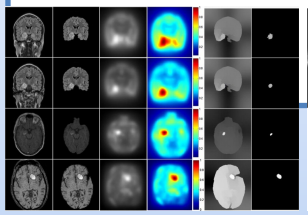
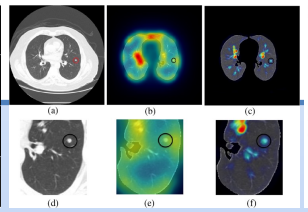
Medical

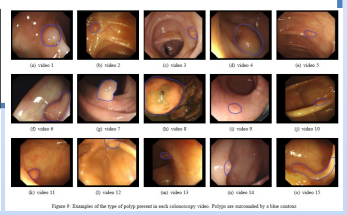
- Medical data prioritization

→ **Detect significant content in medical videos and present it to the expert to save time**

- Efficiently build medical repositories using video summaries

→ **Video summarization allows efficient indexing and searching of medical videos**



Surveillance

- Content prioritization

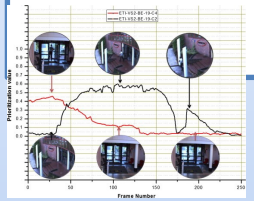
→ **Prioritize surveillance streams based on event importance**

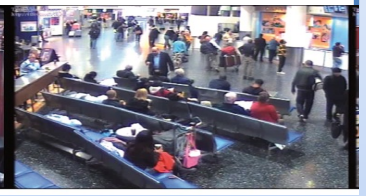
- Save bandwidth

→ **Only transmit significant video data for further processing and storage**

- Interesting events detection

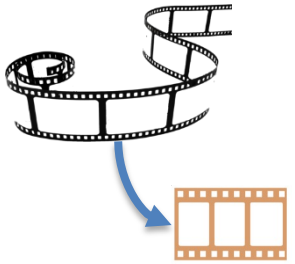
→ **Process high priority contents for interesting event detection**



INDEX

02



II. Video Summarization Methods

Video Summarization Methods classification

No. of Views Based Video Summarization Classification

Video Summarization Techniques

Video Summarization Methods classification

1 Cluster Based Methods

Classification of cluster based methods

- Similar activity based methods
- K-means based methods
- Partitioning based methods
- Spectral based methods

2 Visual Attention Based Methods

Classification of visual attention model based methods

- Motion saliency based methods
- Multi-scale contrast saliency based methods
- Texture saliency based methods
- Object based methods

3 Event Based VS using Deep Features

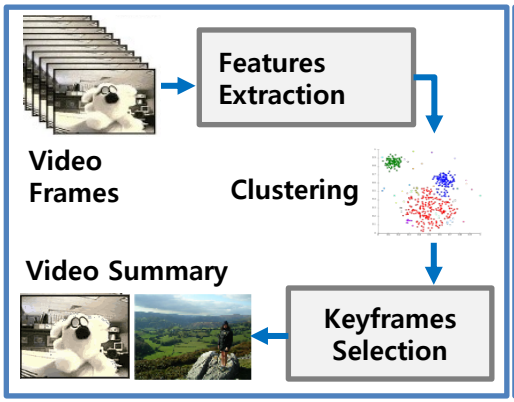
Classification of event based methods

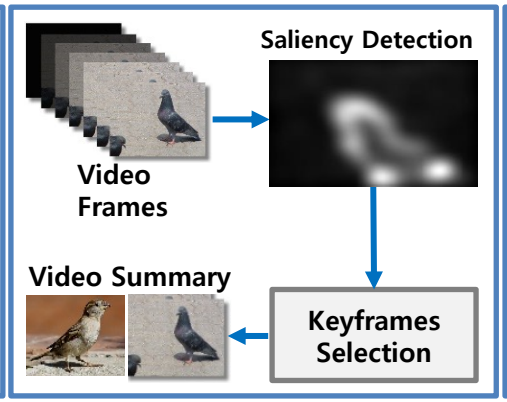
- General event based methods
- Deep features based event-assisted VS methods

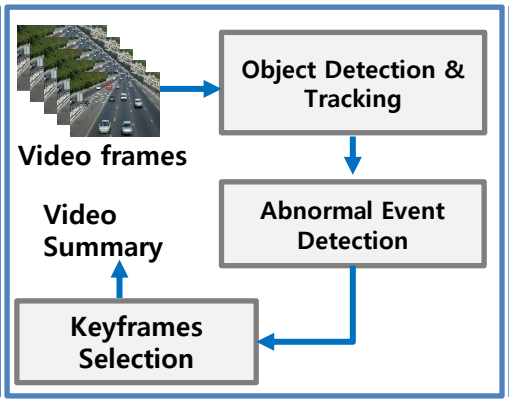
4 Multi-Modality Based Methods

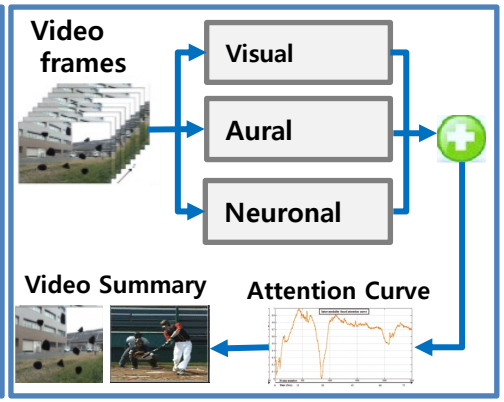
Modality-wise classification

- Visual features based methods
- Audio features based methods
- Hybrid methods









→ Video summarization methods to better understand video contents can generate effective video summaries

No. of Views Based Video Summarization Classification

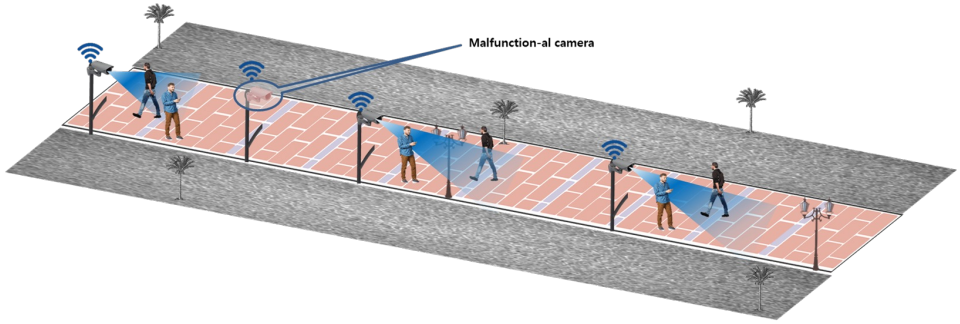
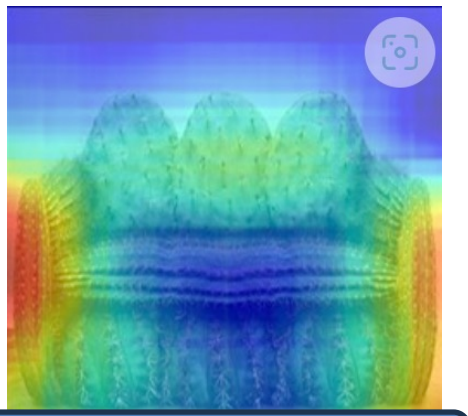
Term	Definition	Visualization
Multi-view overlapping cameras	•These are the multi-view cameras installed in airports, etc. •Multi-view cameras provides broader coverage •Overlapping cameras have high level of intra-view correlations	
Single-view cameras	•Non/least overlapping cameras with no intra-view correlation •Such cameras are comparatively less challenging •Most of streets have single-view cameras with no overlapping.	

Image Memorability

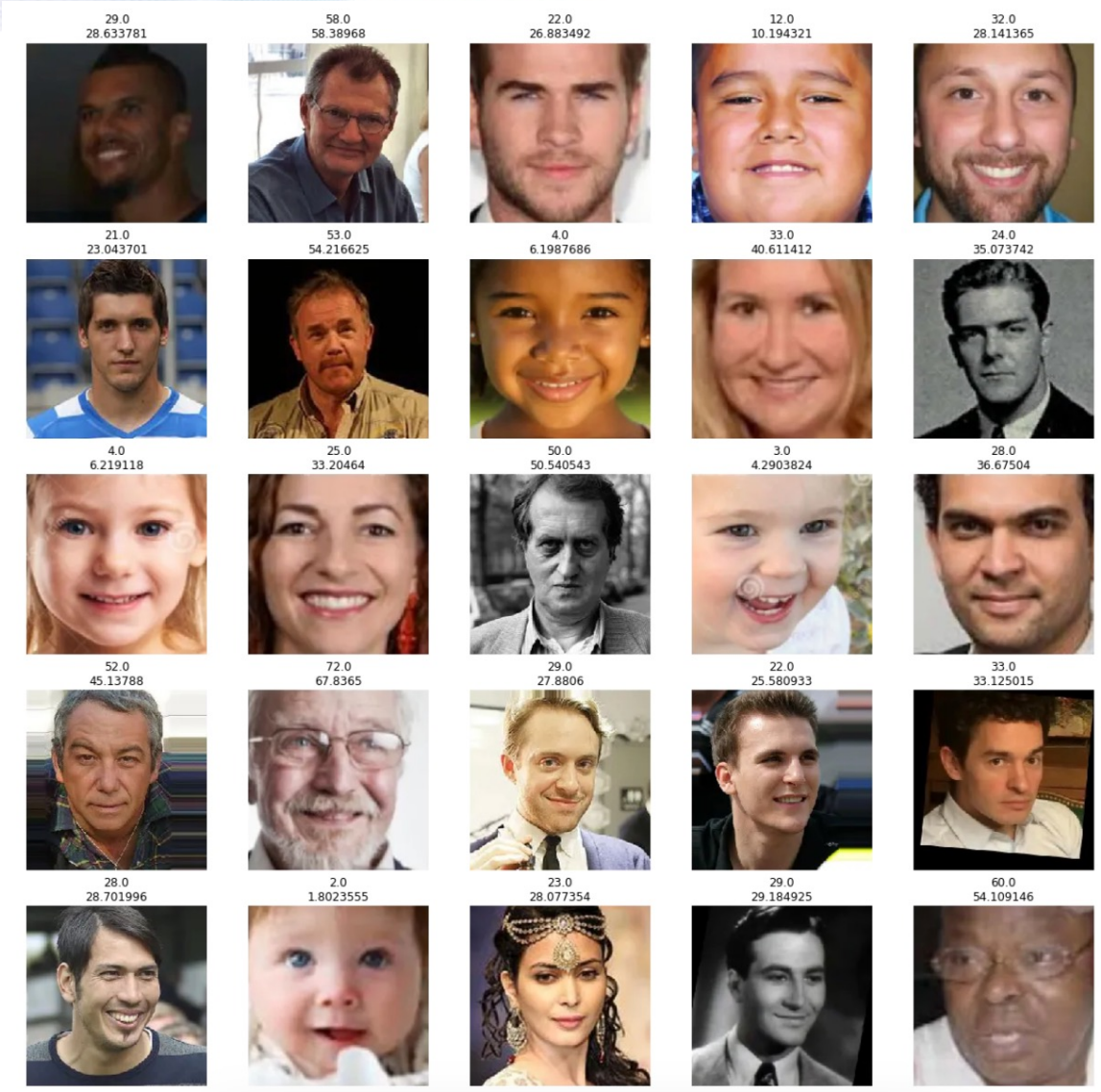


High

Low

Image Regression



Age prediction

Image Memorability as VS Technique?

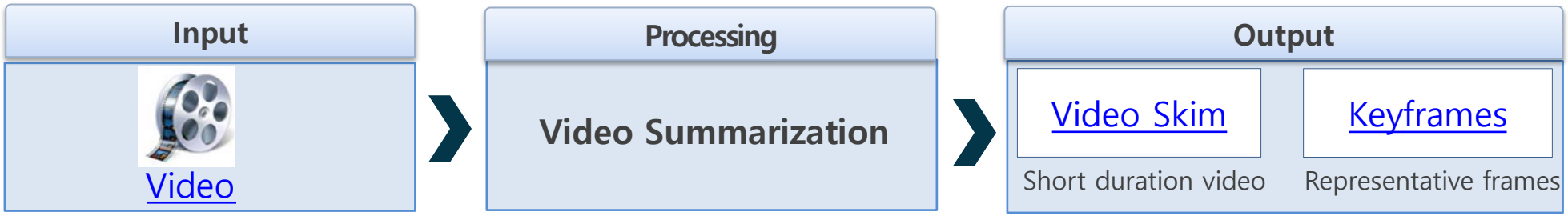
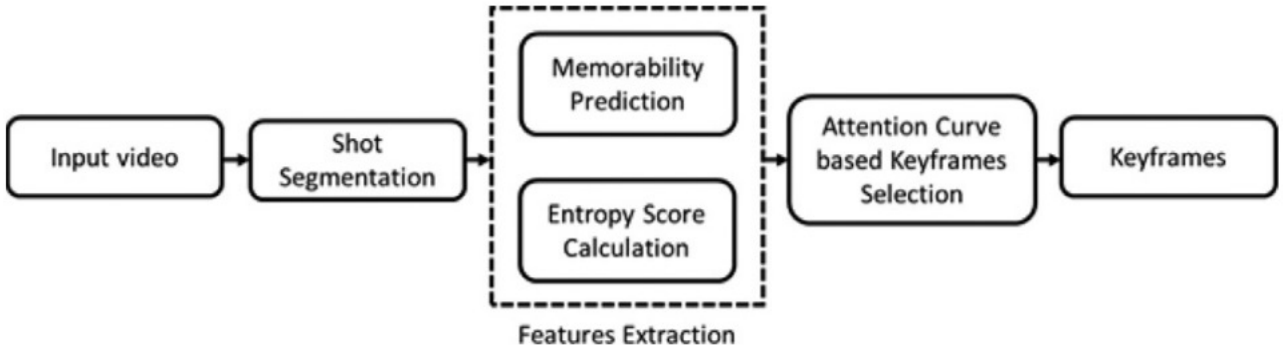
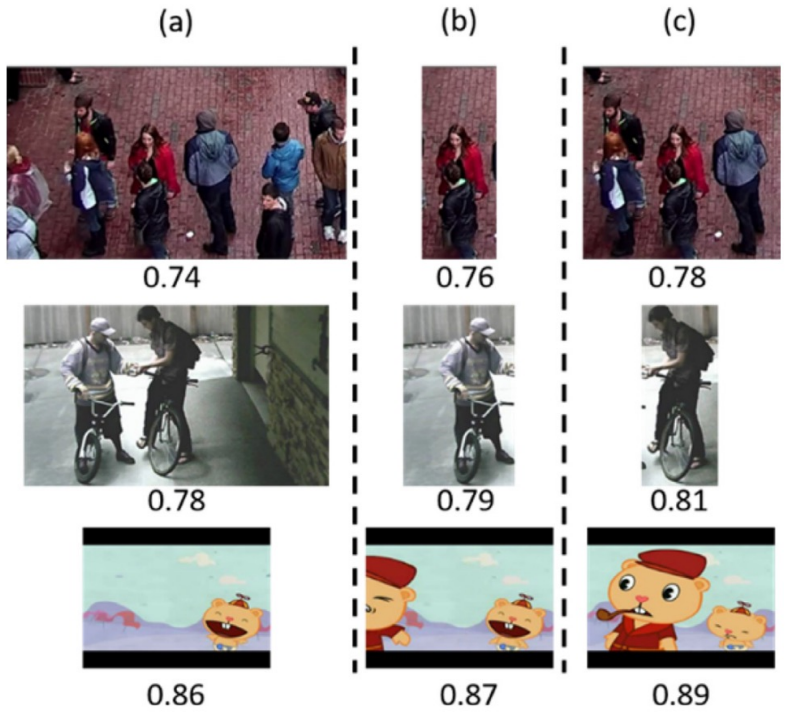


Image memorability?

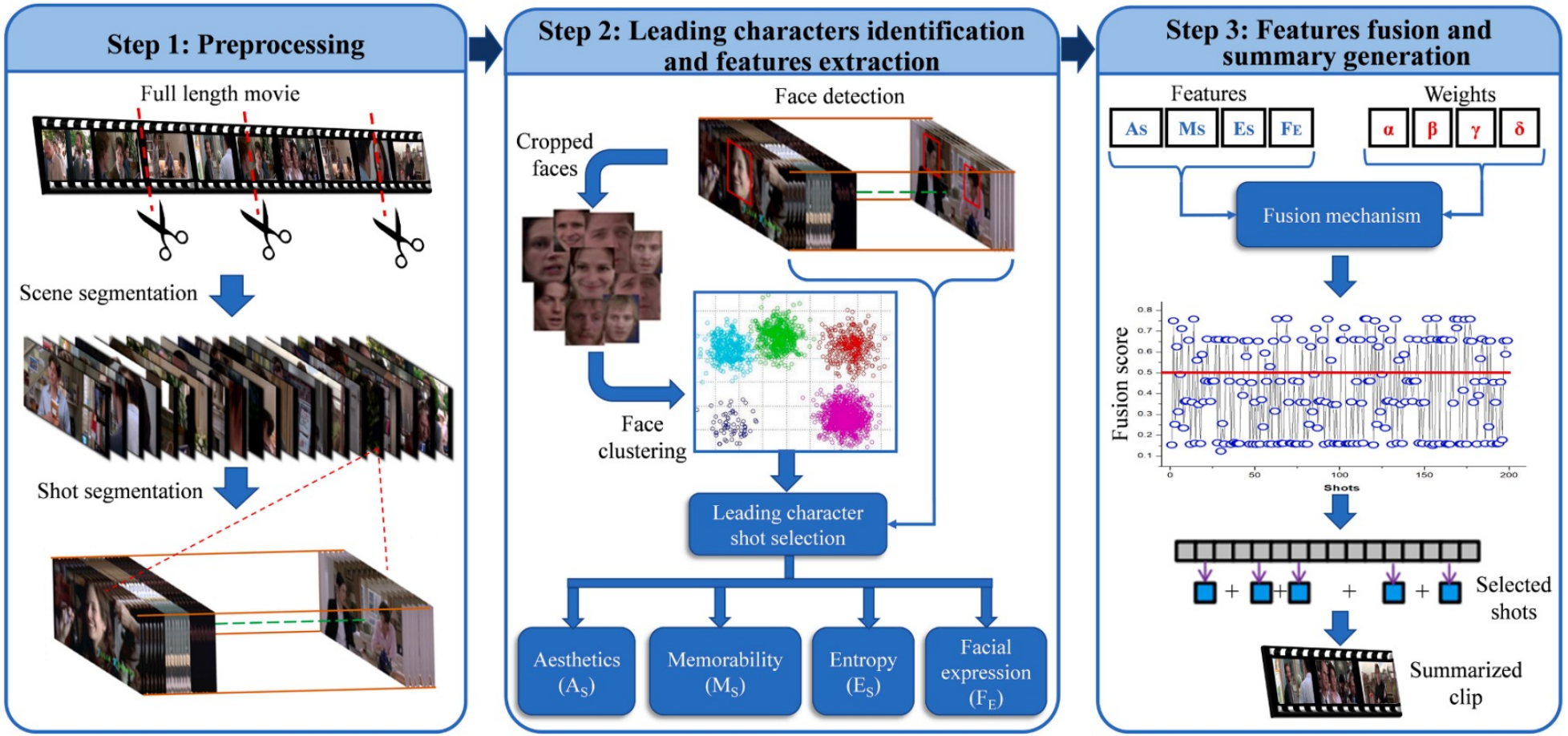
Image Memorability-based Summarization





QuickLook

[QuickLook: Movie summarization using scene-based leading characters with psychological cues fusion]



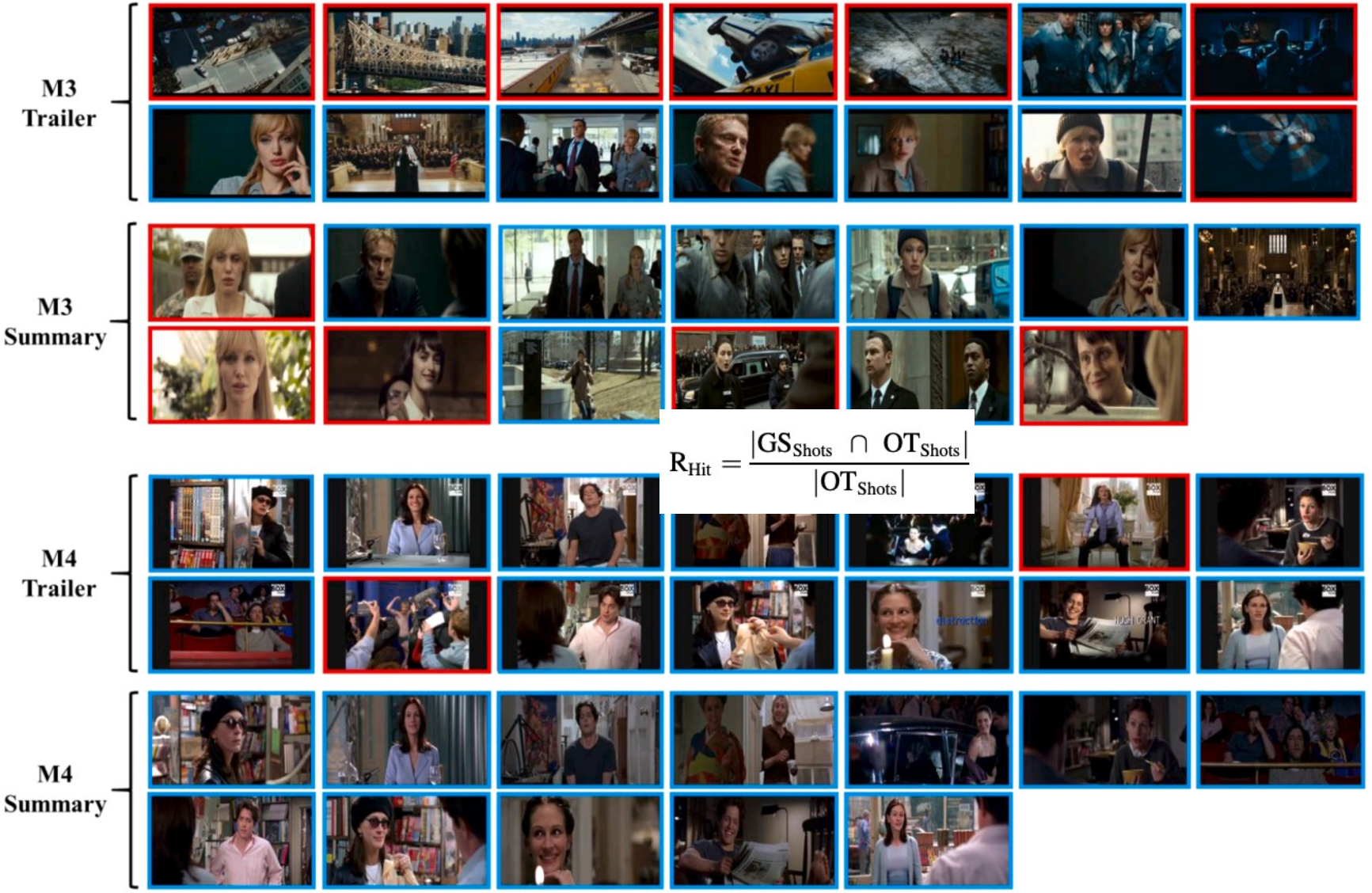
QuickLook

[QuickLook: Movie summarization using scene-based leading characters with psychological cues fusion]

Shot# 19 Shot# 62 Shot# 132 Shot# 177

Shot Number	19	62	132	177
Aesthetics Score	0.79	0.21	0.42	0.89
Memorability Score	0.89	0.39	0.44	0.78
Entropy Score	0.81	0.48	0.51	0.61
Facial Expression	0.7	0.5	0.9	0.9
Fused Score	**0.78**	**0.38**	**0.58**	**0.81**

QuickLook



$$R_{Hit} = \frac{|GS_{Shots} \cap OT_{Shots}|}{|OT_{Shots}|}$$

(b)

3.2. Movie Trailers Generation

3.2 Multi-view video summarization using CNN and bi-directional LSTM

Motivation

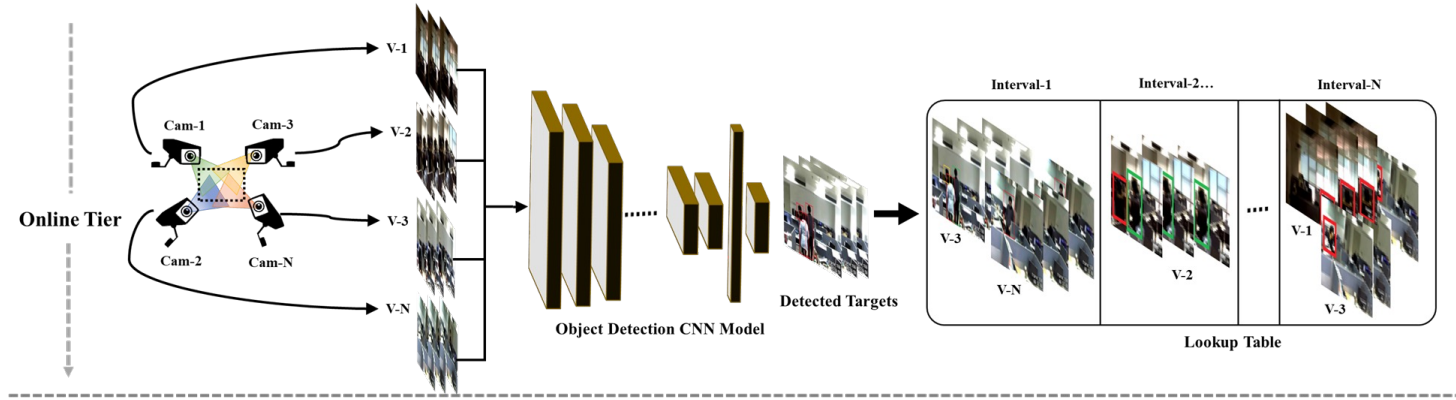
- Single-view surveillance has limited coverage.
- Multi-view video summarization (MVS) is inadequately covered using recent emerging deep learning based methods.

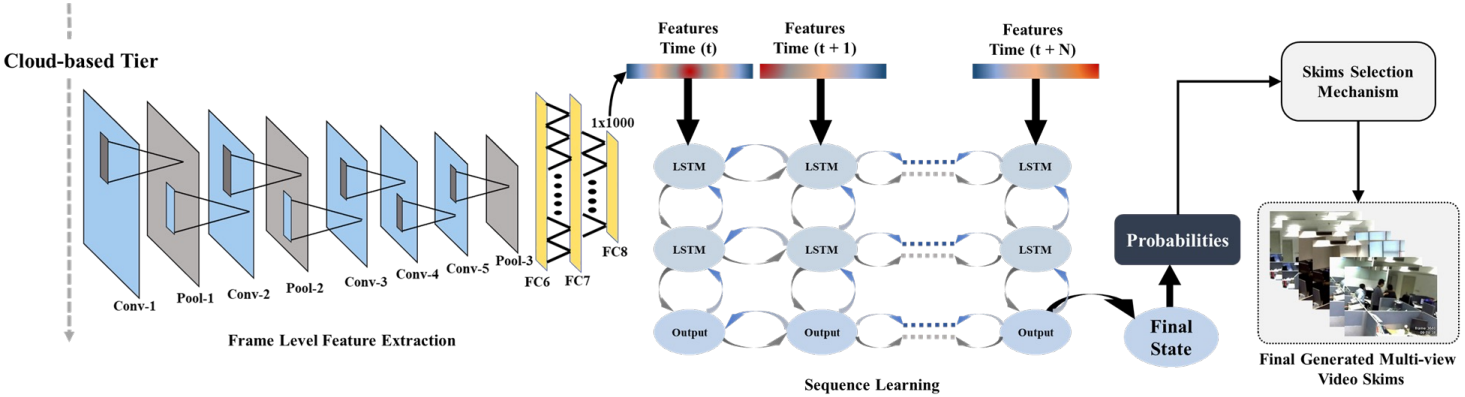
Method description

- Targets appearance based shots segmentation mechanism.
- Deep features extraction from dense layers of CNN architecture.
- Final skims generation via bi-directional LSTM for keyframes selection decision

Results

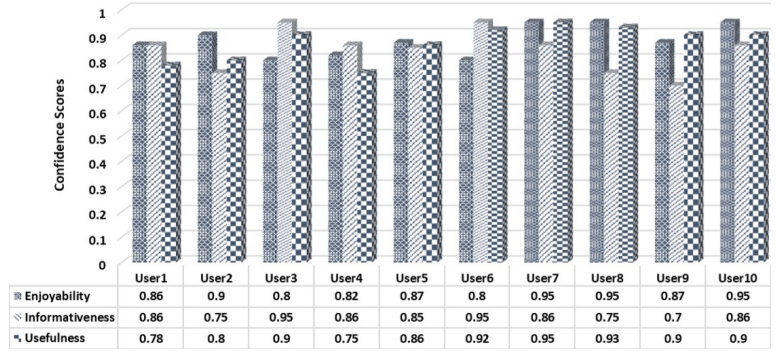
- Informativeness score between 0.7 and 0.95 assigned by participants for generated MVS. 0.1 increase in F1-score





View	Precision	Recall	F1 Score	Event Recall
office-0	0.91	0.88	0.89	0.88
office-1	0.94	0.85	0.89	0.85
office-2	0.94	0.88	0.91	0.88
office-3	0.92	0.85	0.88	0.87

Platform	Processing time (seconds)
Local server/ Personal computer	2048.88
Cloud server/GPU	343.01



INDEX

03

III. Challenges & Future Directions

1. Major Challenges
2. Future Directions

1. Major Challenges

Semantic gap

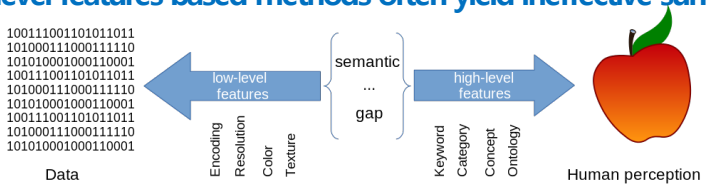
Semantically similar but visually different

- Features fail to describe high level visual semantics

→ **Low-level features-based image difference or visual attention is not adequately effective in representing human interest.**

- Semantic gap leads to ineffective summarization

→ **Low-level features based methods often yield ineffective summaries**



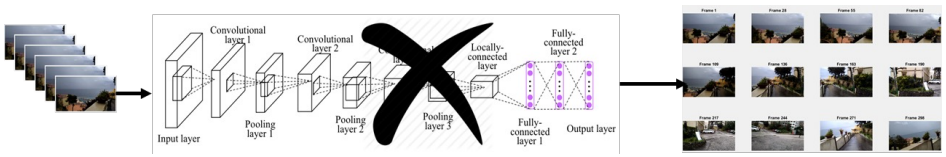
Lack of end-to-end architectures for deep learning

- No standard method of end-to-end learning for VS

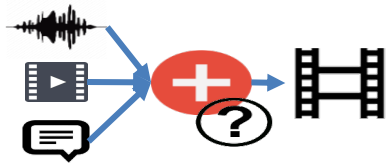
→ **Lack of deep learning methods for efficient video summarization**

- Deep features based approaches

→ **Using deep features improve representation but VS needs human-oriented content representation**



Multi-modality information fusion for VS



- How important is each individual modality to VS?

→ **It is difficult to determine the importance of each modality for VS**

- What other modalities can contribute towards effective VS?

→ **Need to explore other modalities for VS**

Lack of standard methods for performance evaluation

- Lack of standard objective evaluation metrics

→ **Difficult to obtain ground truth due to the subjective nature of VS**

- Subjective evaluation varies from user to user

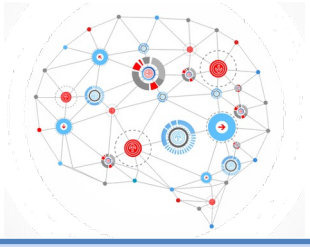
→ **Summary generated according to the interest of one user may not effectively represent the interest of other users**

2. Future Research Directions

Minimize semantic gap using deep learning with big data

- Utilize the power of deep learning and big data

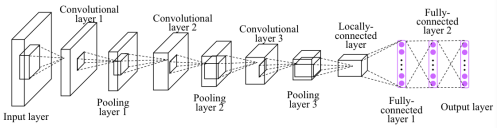
→Train powerful architectures using multimedia big data for semantic video understanding and VS



- End-to-end learning based VS

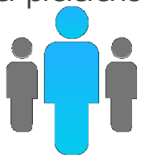
→Deep CNN, LSTM, and deep reinforcement learning presents unlimited opportunities for improvement

→Deep architectures to produce interesting summaries from raw video input needs to be investigated



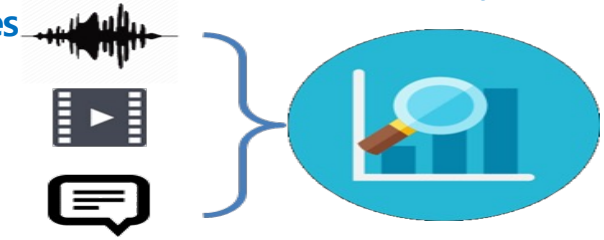
Personalized video summarization

User preferences



- Generate personalized video summaries

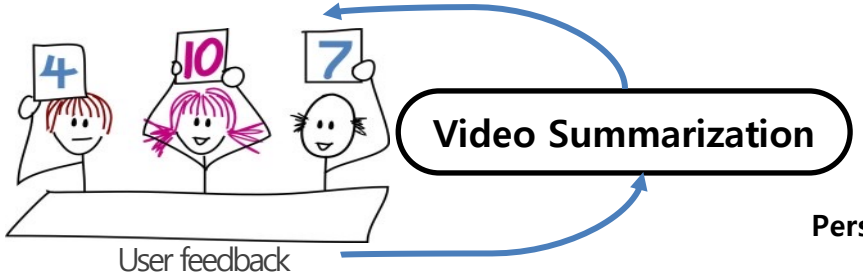
→ Determine the worth of each modality as per user preferences



- Understand user intent, then summarize

→ Grab user preferences and include them into the VS framework

→ Iteratively update summaries based on user feedback



Personalized video summary

Thank you
Q&A