Video Summarization

Dr. I. Mademlis, E. Charalampakis, M. Kaseris, P. Alexoudi, C. Aslanidou, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 3.0

Artificial Intelligence & Information Analysis Lab

Contents

Contents

Artificial Intelligence & Information Analysis Lab

Introduction

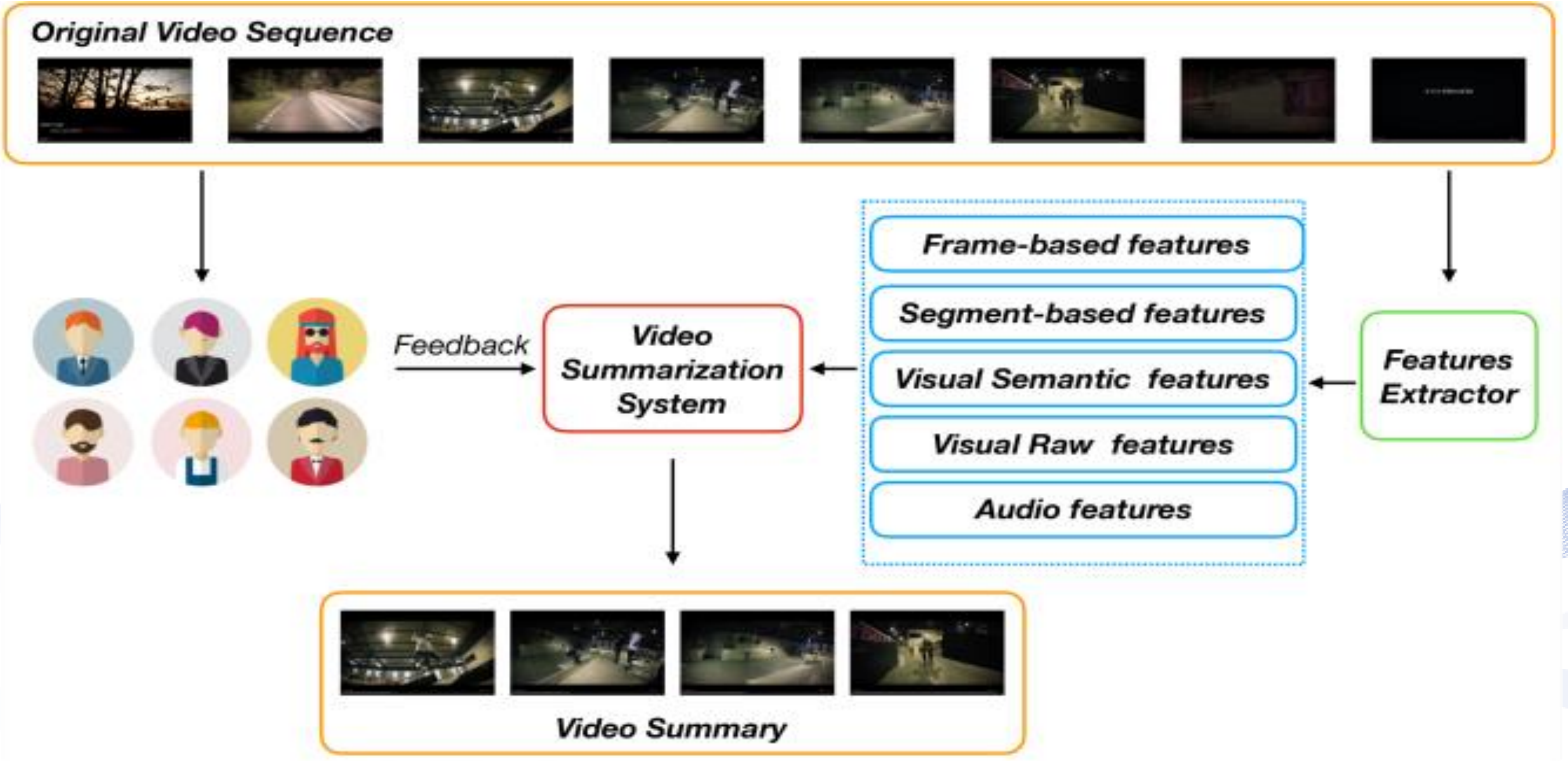
- ***Video summarization*** is the automated construction of a short version of an original full-length video.

- It is necessary in applications where videos are recorded, stored and accessed in abundance.

- Video summarization has various applications in several industries (media, surveillance, WWW, etc.).

- ***Example***: Users would ideally like to browse quickly through large video databases, to get an idea of the content.

Introduction



(Image from Heartbeat Fritz AI)

Introduction

- Video summarization algorithms result in a short ***summary*** of the video.
- The challenge is to automatically select which content will be retained and which will be discarded during the summarization process.
- Only the ***most informative*** and/or interesting parts should be kept.

Video summarization use-cases

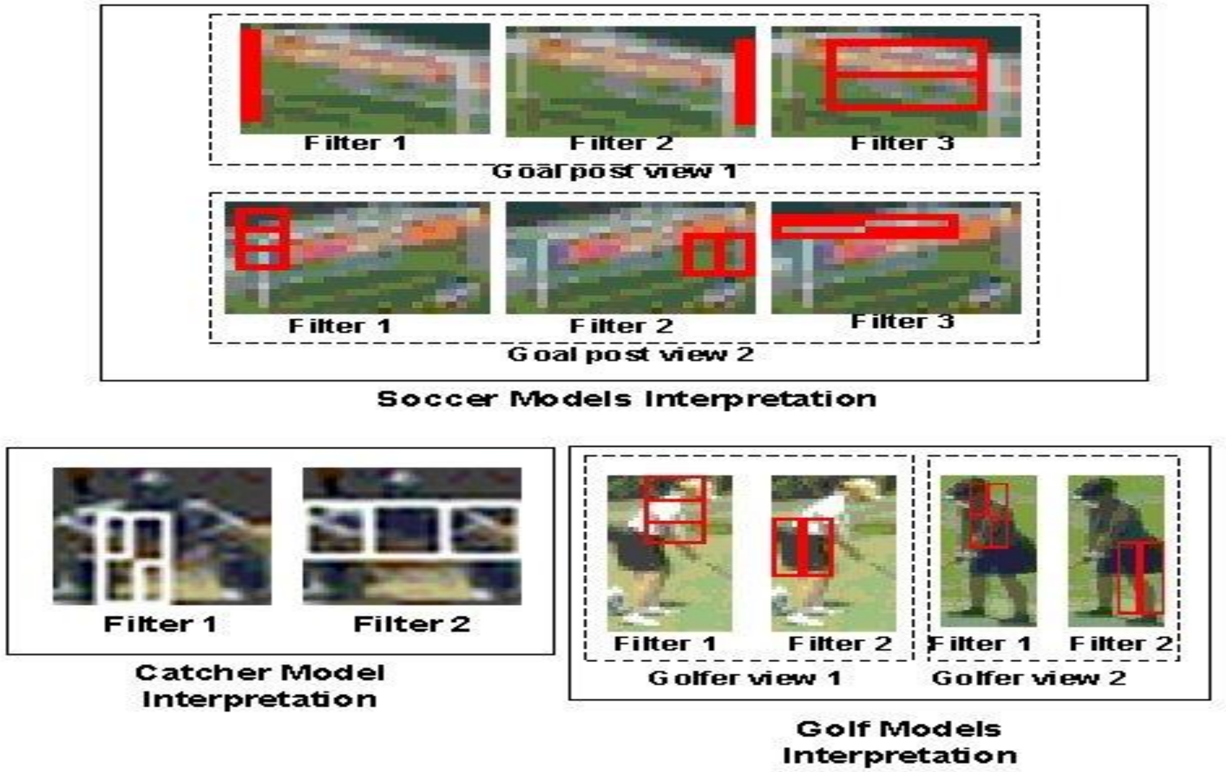
- ***Movie trailers***
- Advertisement creation
- Sports highlights
- ***Anomaly detection in video surveillance***
- Redundancy removal
- Reduction of computational time, storage requirements
- Data visualization
- Search, Retrieval, Recommendation [WOR2020].

Video summarization use-cases

- ***Summarization of personal videos*** [DAR2014].

Video summarization use-cases

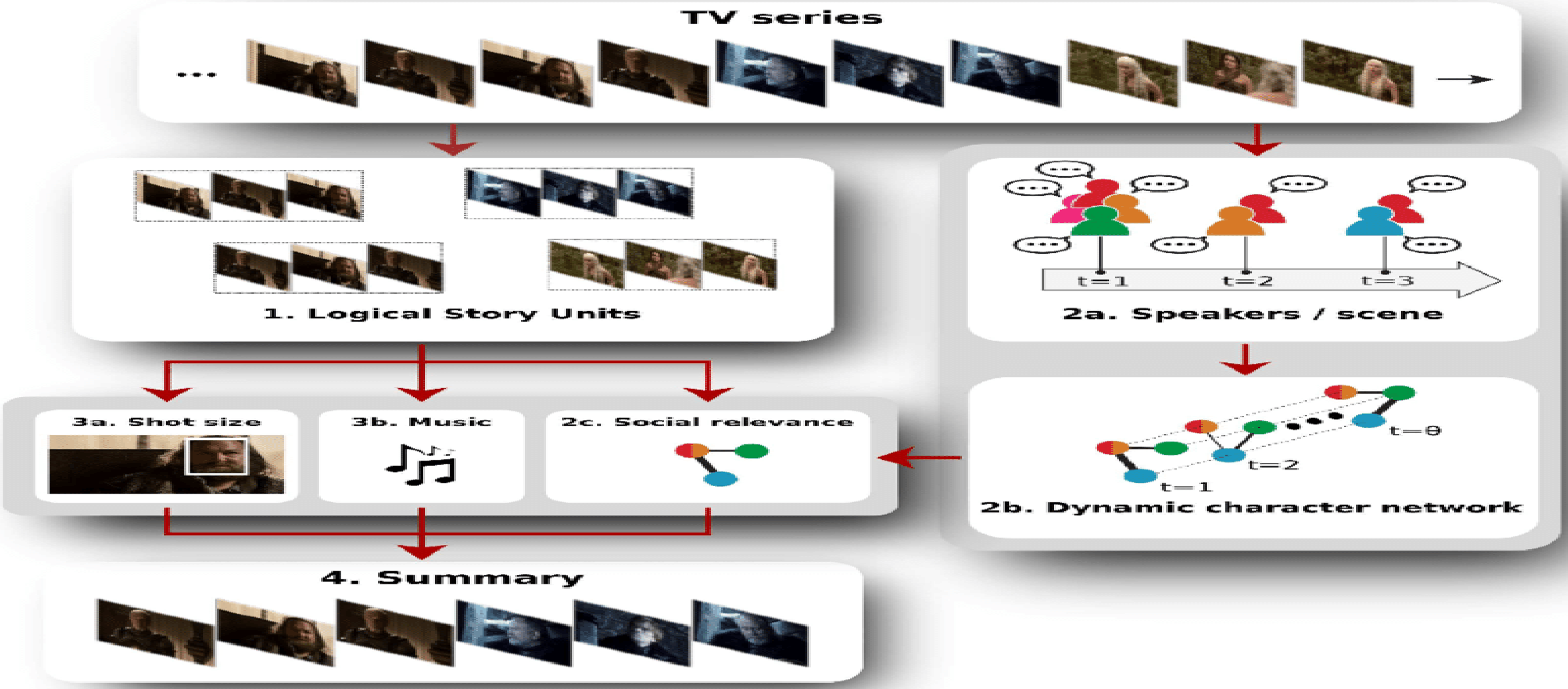
- ***Sports highlights*** [ZHA2006].



Artificial Intelligence & Information Analysis Lab

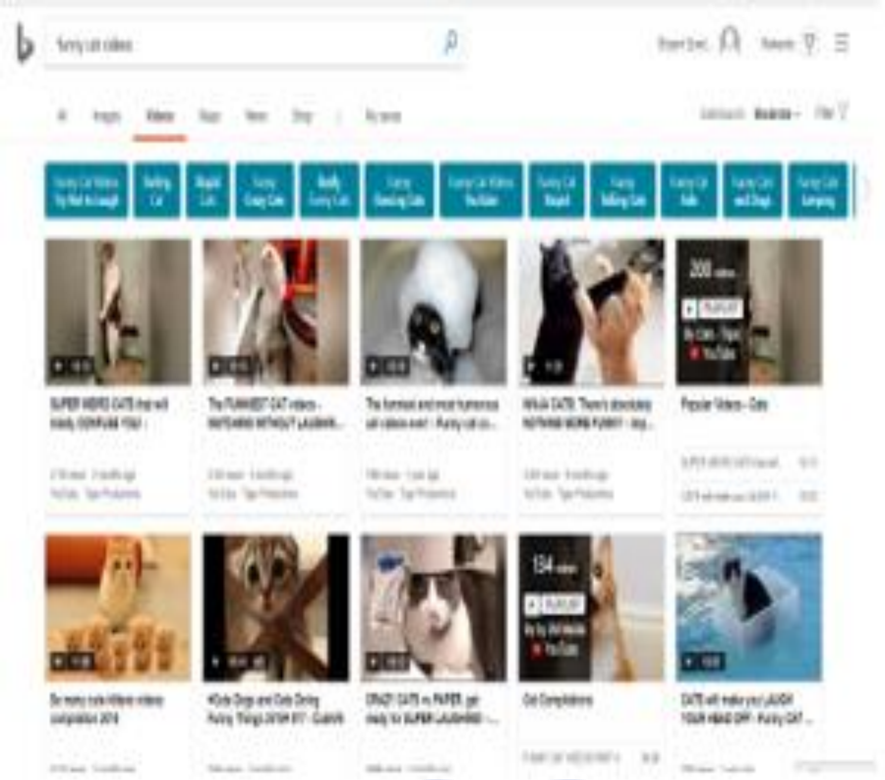
Video summarization use-cases

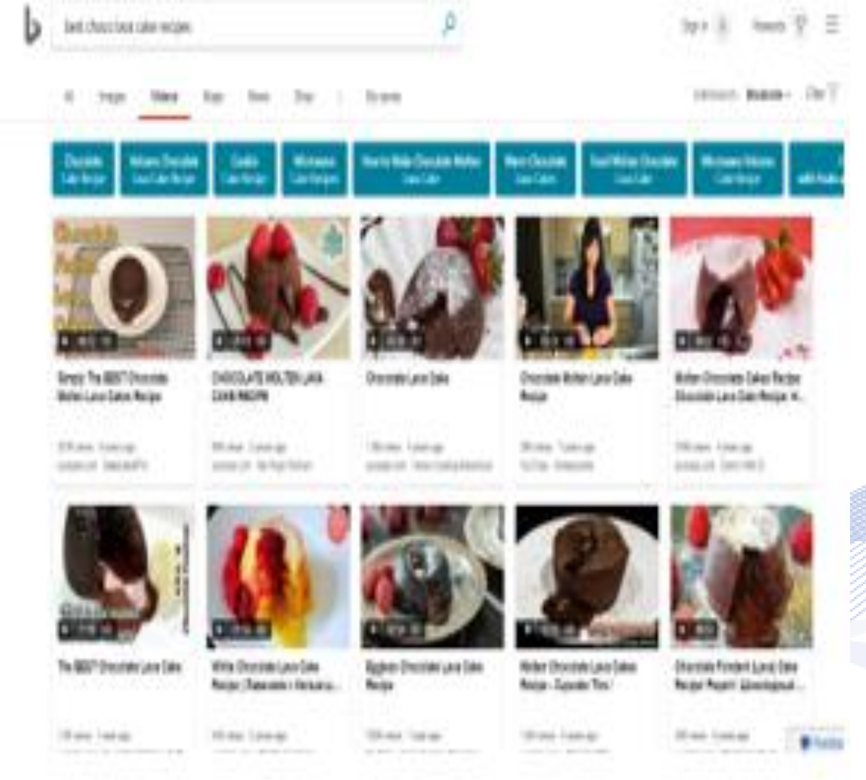
- ***Automatic TV/film trailers*** [BOR2018].



Video summarization use-cases

- ***Video search engines*** [IRI2010].





Video summarization use-cases

- ***Egocentric Video storyboard.***

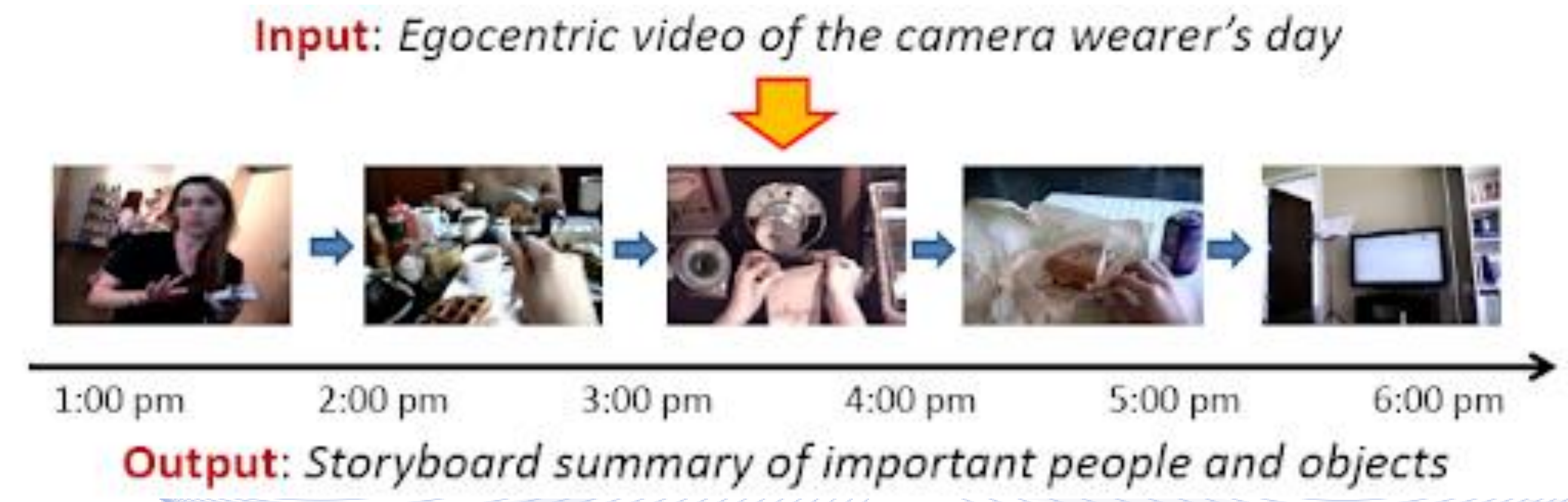


Image from vision.cs.utexas.edu

Video summarization use-cases

- ***Medical Video summarization.***

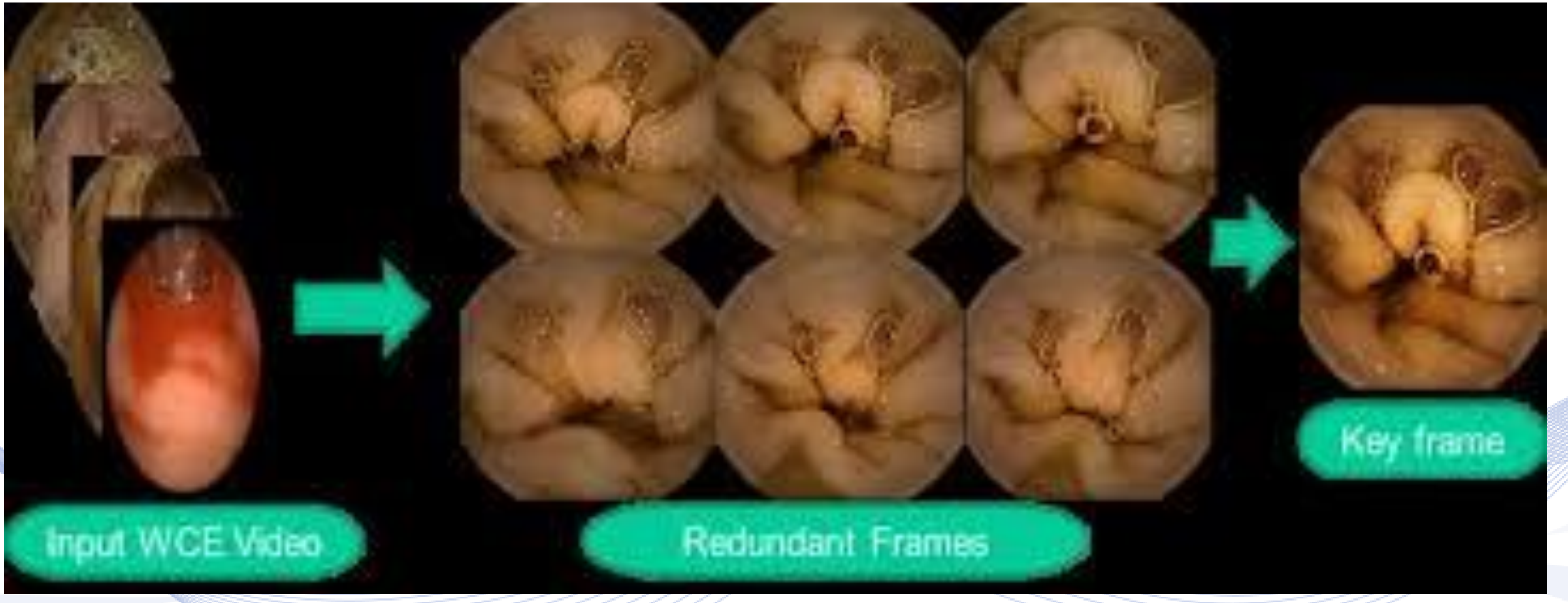


Image from E3S Web of Conferences

Video summarization use-cases

- ***Video inspection***

Artificial Intelligence & Information Analysis Lab

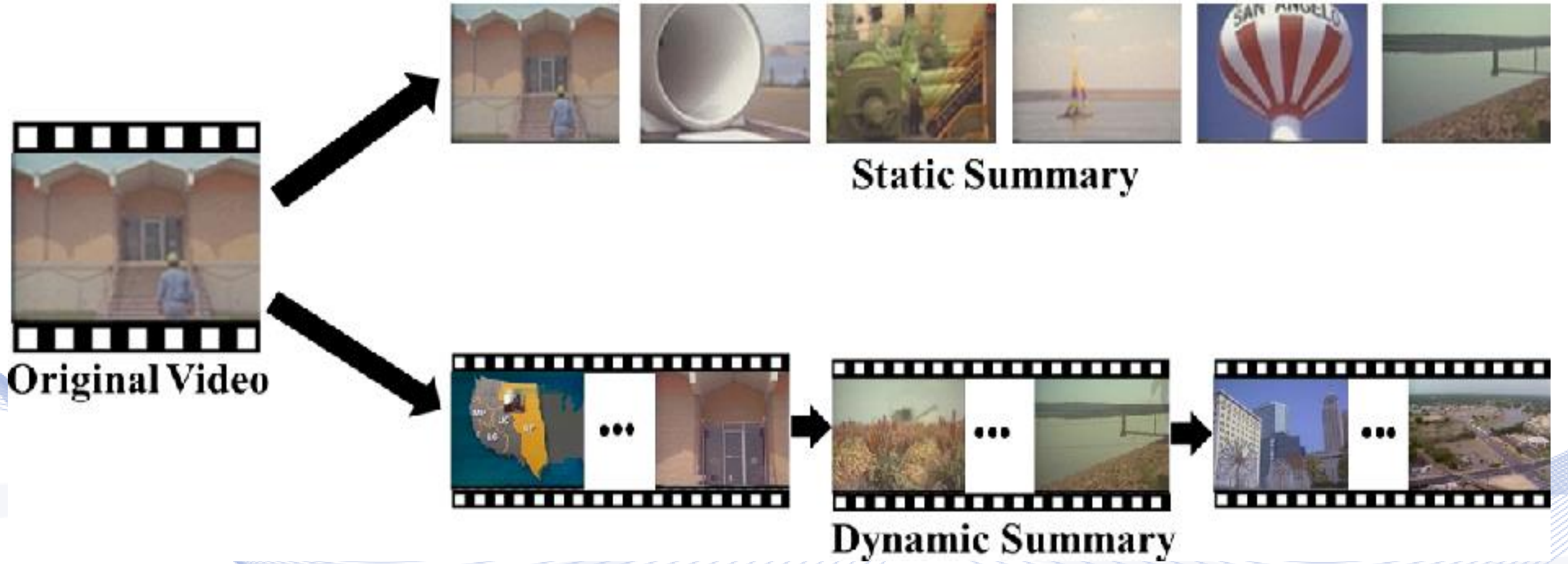
Video summarization use-cases

- ***Natural Disaster Videos***

Video summary types

- There are two main types of video summaries: [MAD2016]
  - ***Static video summaries*** (storyboard/gallery/key-frame set),
  - ***Dynamic video summaries*** (skims/trailers).
- A static summary is a temporally ordered set of selected ***key-frames***.
  - A collection of still images.
- A dynamic summary is a temporally ordered set of selected ***key-segments***.
  - A trailer.

Video summary types



(Image from Semantic Scholar)

Video summarization approaches

- Several video summarization methods have been developed over the years.
- They can be classified into ***four major categories***, based on their properties and characteristics [BUR2020].

Video summarization approaches

- All content selection algorithms for video summarization attempt to identify key-frames/shots/scenes, so that the final summary is:
  - ***Representative*** of the content of the full-length original video,
  - ***Concise*** in length (e.g., the number of key-frames may be 10% of the number of original video frames), and
  - ***Complete***, in the sense that it covers the entire content of the original video (e.g., no sequence of a movie is completely left out of the summary).

Video summarization approaches

- ***Feature-based summarization*** [BUR2020].
  - The original video content is represented by an aggregation of various features.
  - These features may capture properties such as ***visible objects, events, color, motion type***, etc.
  - Feature extraction and aggregation is the most important step.
  - A machine learning method (e.g., clustering) processes these features, in order to select only a subset of the original content.

Video summarization approaches

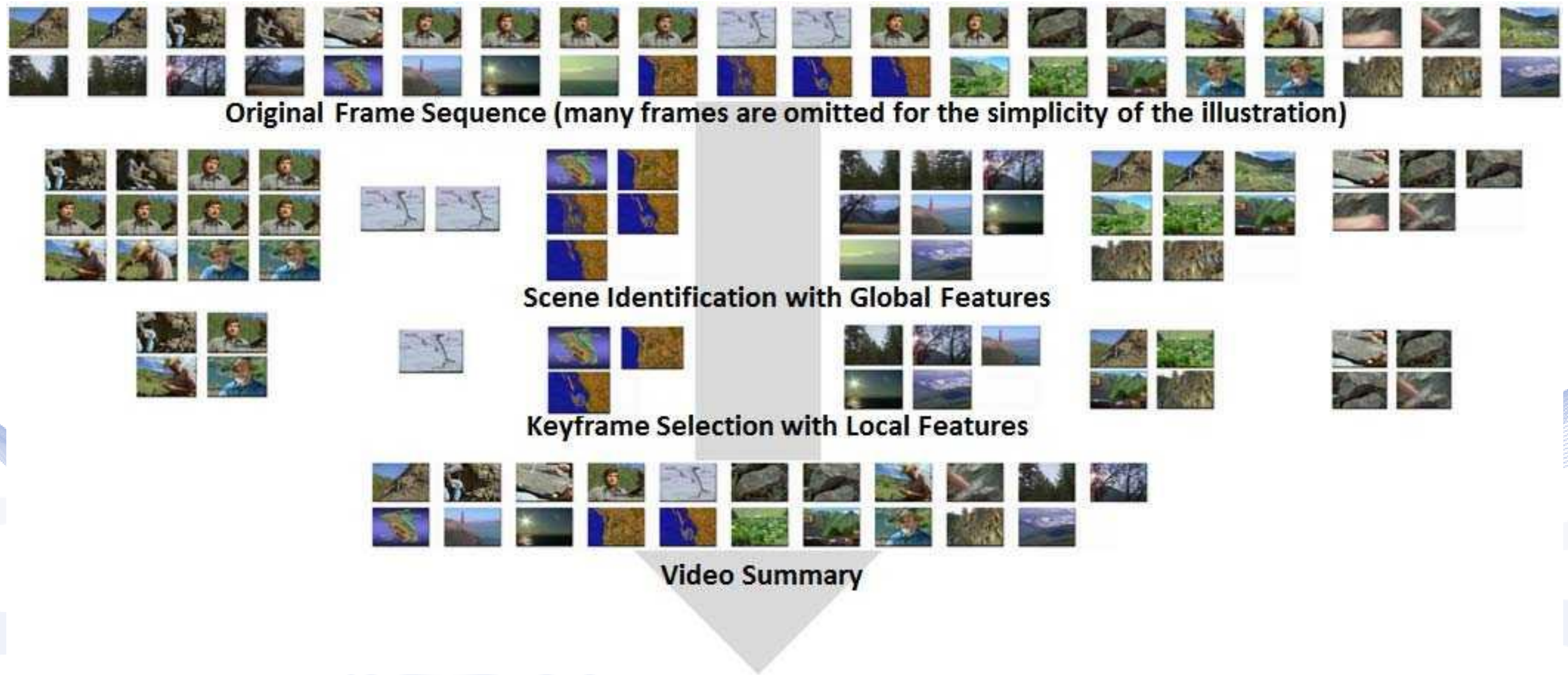
- The selection process may optionally be applied at different levels of detail.
- First, the video is segmented ***into scenes and/or shots***.
- Then, important ***key-scenes*** and/or ***key-shots*** are identified and retained, while the remaining ones are discarded.
- Finally, important ***key-frames*** and/or ***key-segments*** are identified within each of the selected scenes/shots.

Video summarization approaches

Multiple alternative algorithms exist both for temporal video segmentation and for content selection [KAI2012]:

- Clustering of similar video frames into clusters.
- Video change (e.g., shot cut) detection performs temporal video segmentation

Video summarization approaches



Video Summarization with Global and Local Features (Image from ResearchGate).

Video summarization approaches

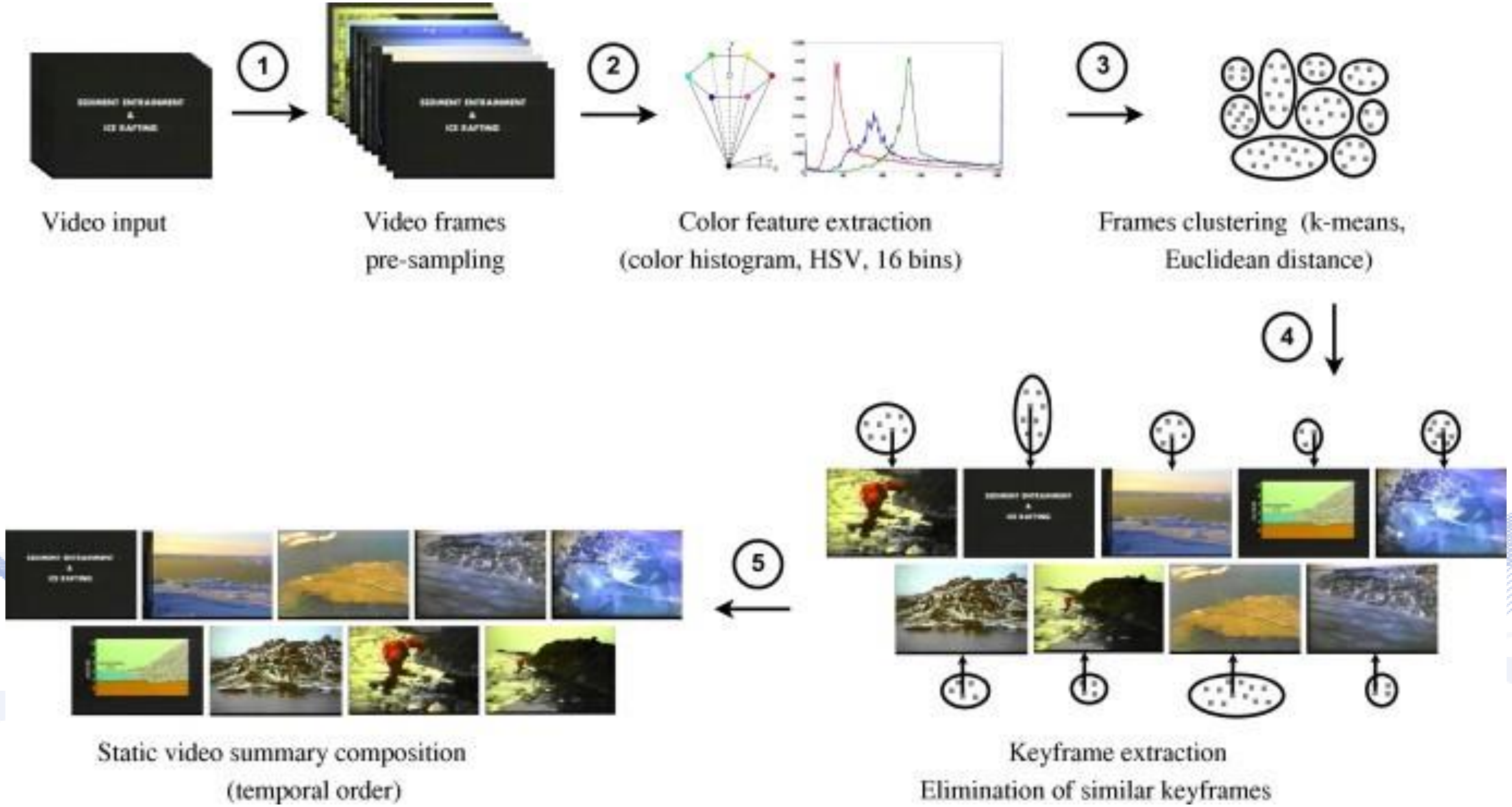


Image from Heartbeat-Fritz AI

Video summarization approaches

- ***Event-based summarization*** [BUR2020].
  - Visually ***abnormal/rare events*** are considered interesting (e.g., a robbery or traffic accident scene in a film).
  - The nature of such events depends on the employed video frame representations:
    - Low-level features expressing perceived motion, colors, etc.
    - Higher-level semantic features expressing visible objects, activities, etc.
  - The selection algorithm retains in the summary only parts of the original video that seem to contain ***abnormal content***.

Video summarization approaches

- ***Event-based video summarization.***

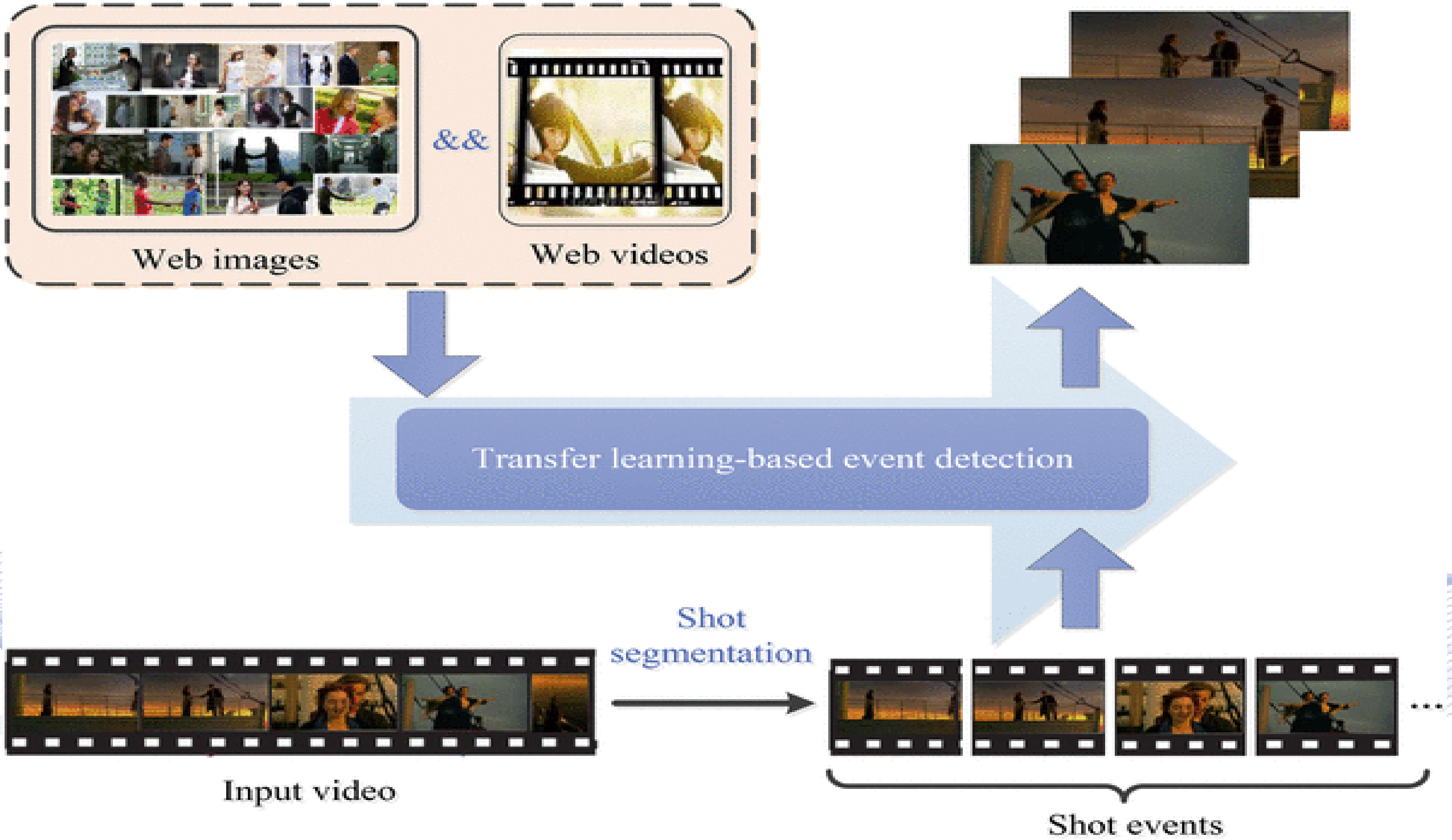


Image from ResearchGate

Video summarization approaches

- ***Object-based summarization*** [BUR2020].
  - There are cases where we are only interested in the parts of the video depicting a specific family of objects (e.g., people).
  - An object detector is required to analyze each scene.
  - Only parts of the original video (frames or segments) containing the desired object(s) are retained in the summary.

Video summarization approaches

- ***Object-based video summarization.***

Image from ScienceDirect

Video summarization approaches

- ***Attention-based summarization*** [BUR2020].
  - There are various ways to identify which parts of an original video hold most of the ***users' interest*** when they view it.
  - The derived summary may only contain key-frames/shots that have been assigned a high attention score.
  - For example, motion ***attention models*** may be employed to measure each shot's interest.

Content selection algorithms

- Various content selection algorithms have been employed for video summarization.
- Video frame/shot/scene ***clustering*** (e.g., K-means) is the simplest approach.

- More sophisticated methods (e.g., ***spectral clustering***) have also been employed.

Artificial Intelligence & Information Analysis Lab

Content selection algorithms

- All video frames are partitioned into ***clusters of similar properties*** and the ***centroid*** of each cluster is retained as a key-frame.
- ***Temporal subsampling*** may be applied before clustering, due to typically high similarities in the appearance of neighboring video frames.
- The exact same process may be applied at a shot or scene level.

Content selection algorithms

- ***Clustering-based Video summarization.***

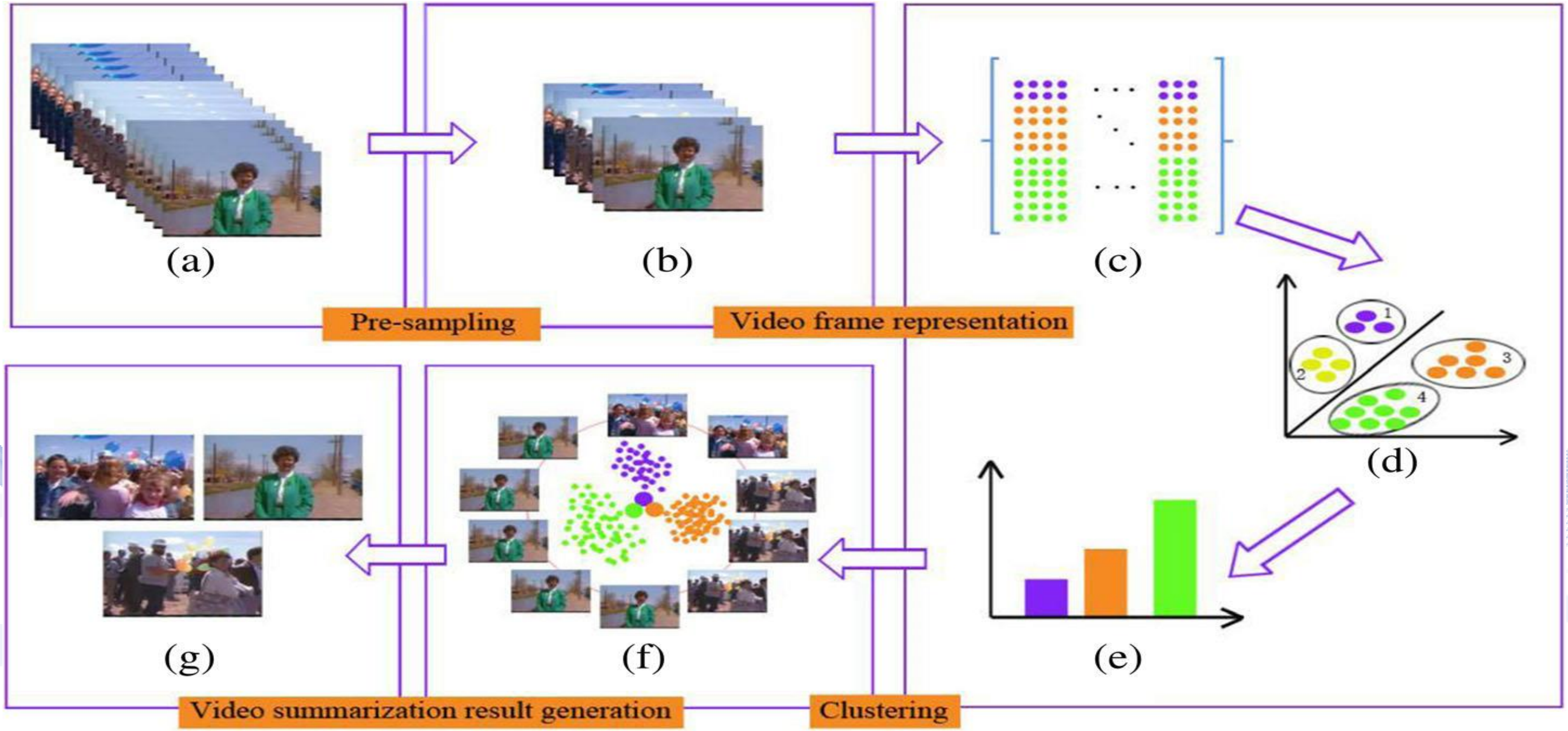


Image from Research Institute for Future Media Computing

Content selection algorithms

- ***Dictionary learning*** is an effective replacement for clustering algorithms.
- The extracted key-frames form a ***dictionary***.
- They should enable ***optimal reconstruction of the original video*** from the selected dictionary.
- Thus, the video summary is framed as the set of key-frames that can linearly reconstruct the full-length video in an algebraic sense [MAD2018].

Content selection algorithms

- ***Sparse dictionary learning.***

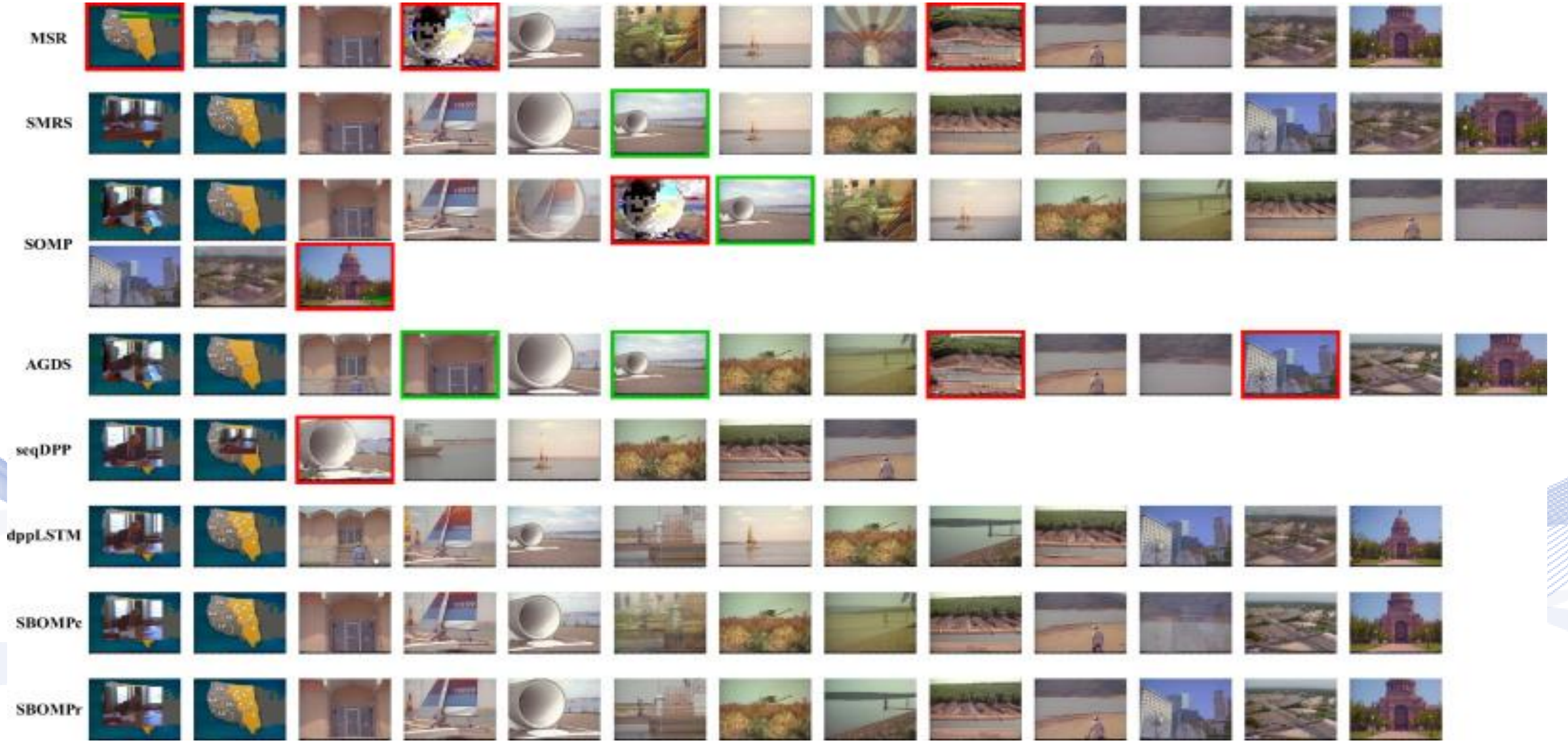


Image from sciencedirect.com

Artificial Intelligence & Information Analysis Lab

Content selection algorithms

- Both clustering and dictionary learning are ***unsupervised learning*** approaches: no ground-truth summaries are required.
- The following approaches have also been proposed:
  - Reinforcement learning [WOR2020] or
  - supervised learning methods [DIN2019] .
- ***Supervised video summarization*** requires training of machine learning model using a manually annotated training dataset.
- The annotation may be an importance score assigned per video frame.

Content selection algorithms

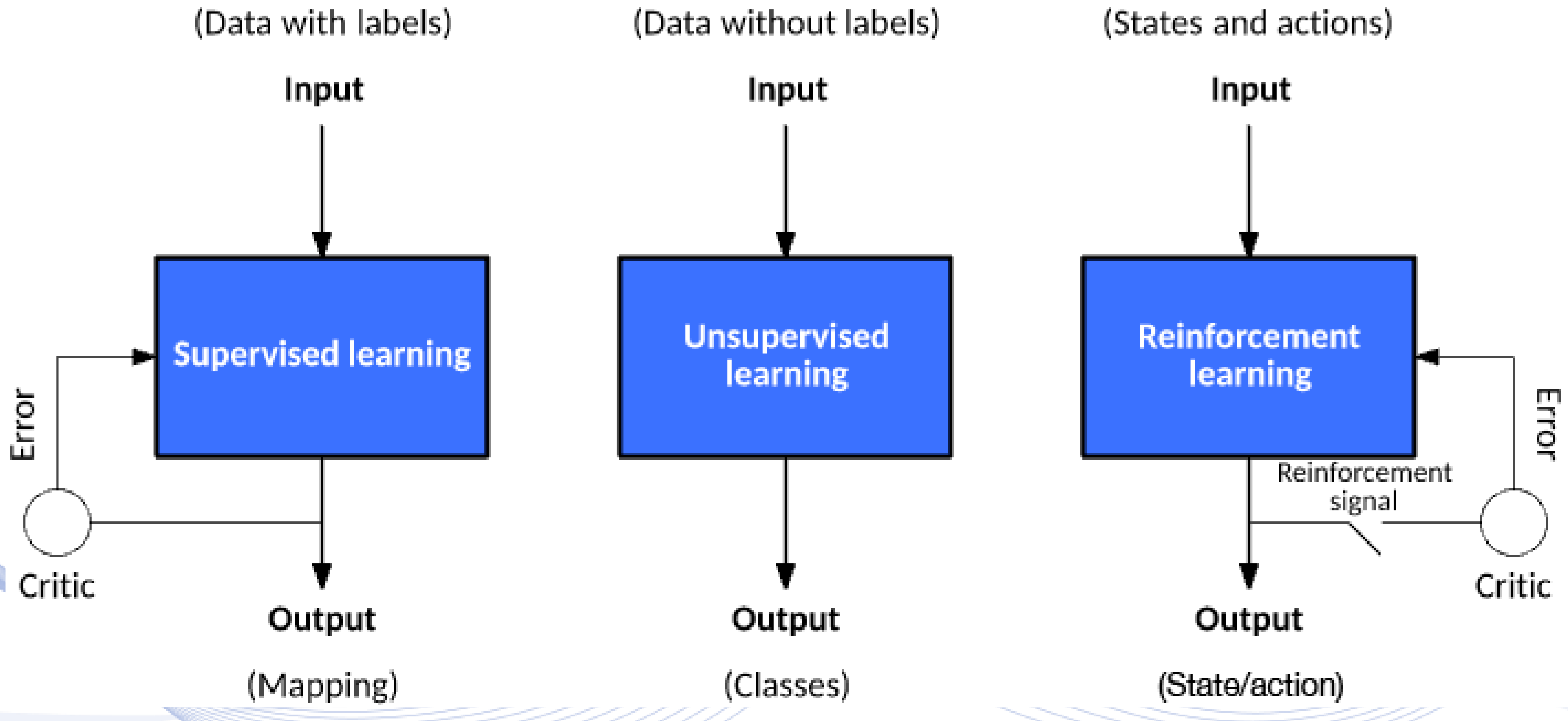


Image from IBM Developer

Content selection algorithms

- The standard supervised approach has several disadvantages.
- ***Manual video annotation is quite expensive***, difficult and costly, especially if done at a per-frame level.
- Importance scores are quite subjective.
- The trained model may only perform well in test videos resembling the training dataset.

Video Summarization with Deep Neural Networks

- In recent years, ***Deep Neural Networks*** (DNNs) have been employed for video summarization in various ways.
- The simplest approach is to exploit semantic video frame representations derived from pre-trained Convolutional Neural Networks (CNNs), as inputs to a traditional content selection algorithm.

Video Summarization with Deep Neural Networks

- A more sophisticated approach is to train a DNN under a supervised learning framework ***to directly regress an importance score*** for each original video frame.

- During the test stage, any video frame which is assigned a score larger than a threshold can be selected as a key-frame.

- This approach has all the disadvantages of supervised summarization.

Video Summarization with Deep Neural Networks

- Various deep neural architectures may be combined in a composite DNN for video summarization. For example:
  - ***Convolutional Neural Networks*** (CNNs)
  - ***Transformers***
  - ***3D CNNs***
  - ***Long Short-Term Memory Networks*** (LSTMs)
  - ***Generative Adversarial Networks*** (GANs).

GANs for unsupervised video summarization

- ***GANs combined with LSTMs*** have recently been employed for unsupervised video summarization, using an end-to-end trainable DNN architecture.

- GANs are generative models which learn the distribution of the training data. They are composed of a Generator and a Discriminator involved in a minimax game.
  - The ***Generator*** learns to generate content that the ***Discriminator*** mistakes for real.
  - After training, the Discriminator may be discarded.

GANs for unsupervised video summarization

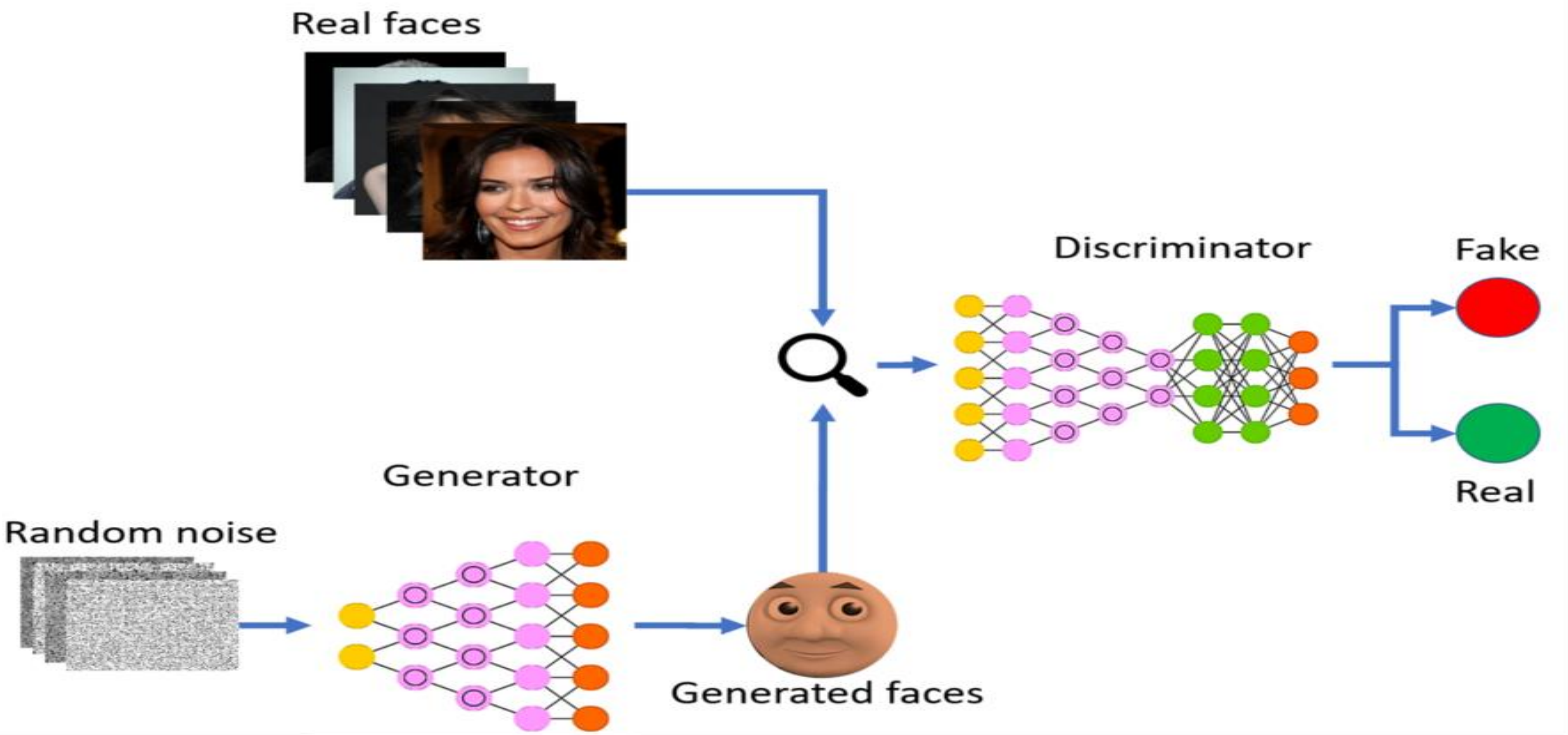


Image from LaptrinhX

GANs for unsupervised video summarization

Examples of fake faces

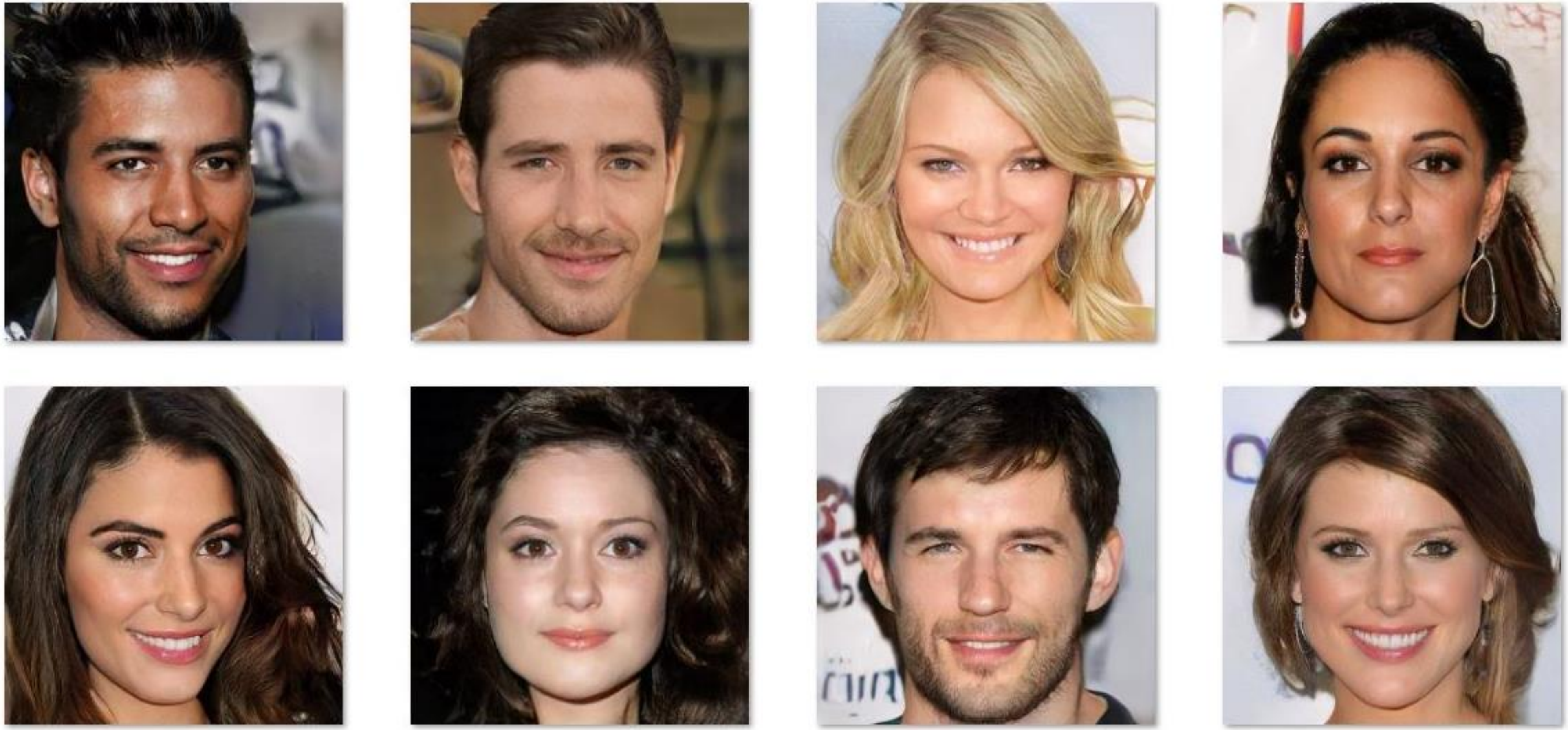


Image from datadriveninvestor.com

GANs for unsupervised video summarization

- CNNs (and possibly LSTM) are used for spatiotemporal video feature extraction.
- Video frames are selected to be included in the summary.
- The Generator tries to generate the video from its summary.
- The Discriminator tries to understand if the generated video is close enough to the original.

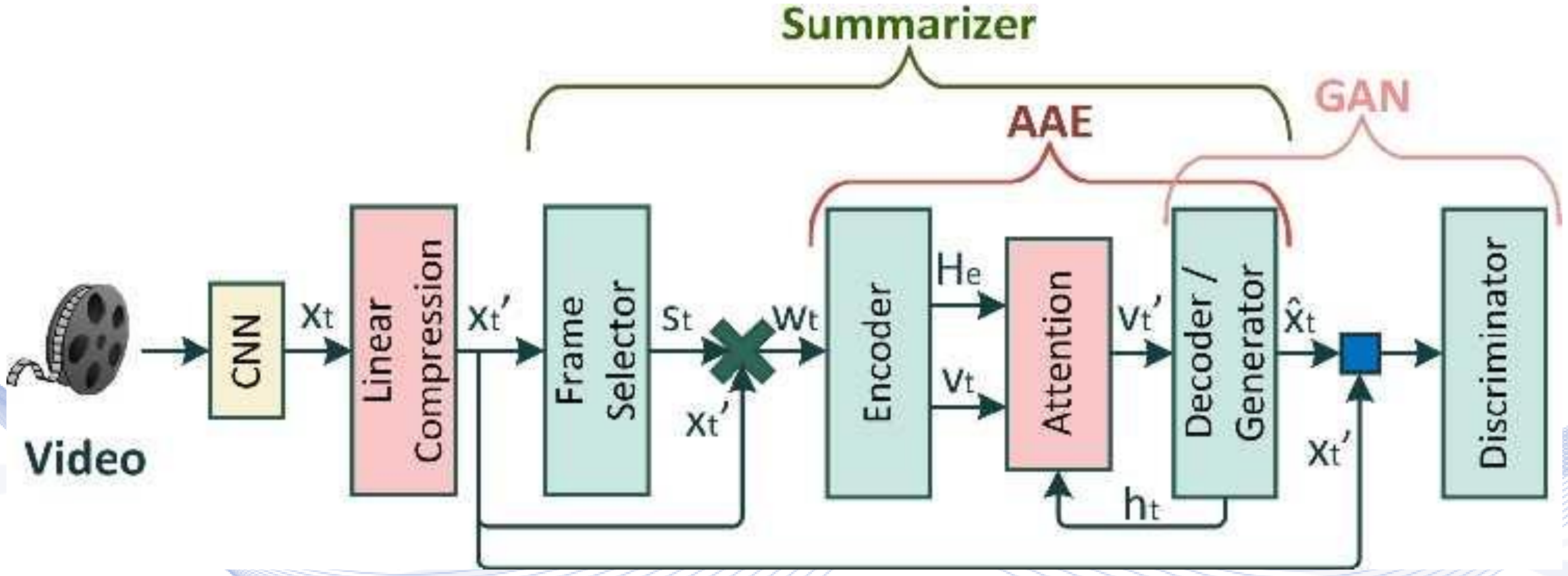
GANs for unsupervised video summarization

- ***SUM-GAN-AAE*** [METS2020].

- ***Dilated Temporal Relational Adversarial Network*** for frame-level video summarization [DIN2019].

- ***Cycle-SUM***: Cycle-consistent Adversarial LSTM Networks for Unsupervised Video Summarization (Video Trailer) [PIN2019].

SUM-GAN-AAE



The architecture of SUM GAN-AAE (Image from [METS2020])

SUM-GAN-AAE

SUM-GAN-AAE is a modification of SUM-GAN [MAH2017].

- The network architecture consists in a ***Summarizer subnetwork***, which acts as a Generator, and a ***Discriminator subnetwork***.
- The Summarizer is a pipeline of three smaller subnetworks:
  - Frame Selector, Encoder, Decoder.
- All subnetworks are LSTMs.
- After training, only the ***Frame Selector*** is required.

SUM-GAN-AAE

- The Frame Selector receives sequentially as input the original video frame representations.

- For each input video frame, it estimates and outputs an ***importance score.***

- The original video frame representations and the importance scores are multiplied.

SUM-GAN-AAE

- The ***Encoder*** is sequentially fed the above products and produces a ***fixed-length representation for the entire video***.

- The representation produced by the Encoder is fed to the Decoder, which is equipped with an ***attention mechanism***.

- The ***Decoder*** is trained to sequentially output the original video frames.

- The Encoder-Decoder and the attention module jointly constitute the ***Attention Autoencoder subnetwork*** (***AAE***).

SUM-GAN-AAE

- Both the original and the reconstructed video frame representations are then sequentially passed to the ***Discriminator***, whose task is to determine whether each sequence is “real” (original) or “fake” (summary-based reconstruction).

- The Frame Selector and the AAE jointly constitute the ***Summarizer***, which is trained to confuse the Discriminator.

  - This forces the Frame Selector to learn how to extract representative key-frames, jointly capable of accurately reconstructing the full-length video.

SUM-GAN-AAE

- $\mathbf{X} \in \mathbb{R}^{M \times N}$: The input video data matrix.
- Each column $\mathbf{x}_i \in \mathbb{R}^M$ of the matrix $\mathbf{X}$, is the feature representation of the i-th video frame.
- The baseline summarization architecture includes:
  - An LSTM-based ***Frame Selector*** S parameterized by weights $\mathbf{w}_s$.
  - An LSTM-based ***Encoder*** E parameterized by weights $\mathbf{w}_e$.
  - An LSTM-based ***Decoder*** D parameterized by weights $\mathbf{w}_d$.
  - An LSTM-based ***Discriminator*** (binary classifier) C parameterized by weights $\mathbf{w}_c$.

Artificial Intelligence & Information Analysis Lab

SUM-GAN-AAE

- S is fed $\mathbf{x}_i$ as input and outputs a corresponding ***scalar importance factor*** $s_i \in [0, 1]$.
- The product $s_i\mathbf{x}_i$ is fed to E resulting in a ***state vector*** $\mathbf{e} \in \mathbb{R}^H$ encoding the summary.
- Subsequently, $\mathbf{e}$ is fed to D which attempts to reconstruct the original $\mathbf{X}$, by outputting a reconstructed $\hat{\mathbf{x}}_i \in \mathbb{R}^M$, $1 \leq i \leq N$.
- Finally, the video reconstruction $\hat{\mathbf{X}}$ is forwarded to the Discriminator C as a “fake” training example, while the original video $\mathbf{X}$ is used as a “real” training example.

Artificial Intelligence & Information Analysis Lab

SUM-GAN-AAE

- The following loss functions are employed during training:
- ***Reconstruction loss***:

$$\mathcal{L}_{recon} = \left\| \phi(\mathbf{X}) - \phi(\hat{\mathbf{X}}) \right\|_2^2,$$

  - $\phi(\mathbf{X})$ is the last hidden LSTM state, when it is fed $\mathbf{X}$ as input.
  - $\phi(\hat{\mathbf{X}})$ the corresponding hidden LSTM state when C is fed $\hat{\mathbf{X}}$.
  - $\mathcal{L}_{recon}$ is used to update $\mathbf{w}_s, \mathbf{w}_e, \mathbf{w}_d$.

SUM-GAN-AAE

- ***Original video loss***:

$$\mathcal{L}_{orig} = \left(1 - C(\mathbf{X})\right)^2.$$

- It is the MSE between the original video label (i.e., 1) and the discriminator output (in $[0,1]$) when C is fed $\mathbf{X}$ as input.
- $\mathcal{L}_{orig}$ updates $\mathbf{w}_c$.
- ***Summary loss***:

$$\mathcal{L}_{sum} = \left(C(\hat{\mathbf{X}})\right)^2$$

- is the MSE between the summary label (i.e., 0) and the computed probability when C is fed $\hat{\mathbf{X}}$ as input.
- $\mathcal{L}_{sum}$ updates $\mathbf{w}_c$.

SUM-GAN-AAE

- ***Generator loss***:

$$\mathcal{L}_{gen} = \left(1 - C(\widehat{\mathbf{X}})\right)^2.$$

- It is the MSE between the original video label (i.e., 1) and the discriminator output, when C is fed $\widehat{\mathbf{X}}$ as input. $\mathcal{L}_{gen}$ updates the Decoder parameters $\mathbf{w}_d$.
- ***Sparsity Loss*** pushes the Selector towards assigning high importance (i.e., key-frame status probability) to a ***small percentage*** of the total number of original video frames, defined by a scalar hyperparameter $\sigma \in [0,1]$:

$$\mathcal{L}_{sparsity} = \left\|\frac{1}{N}\sum_{t=1}^{N} s_t - \sigma\right\|_2.$$

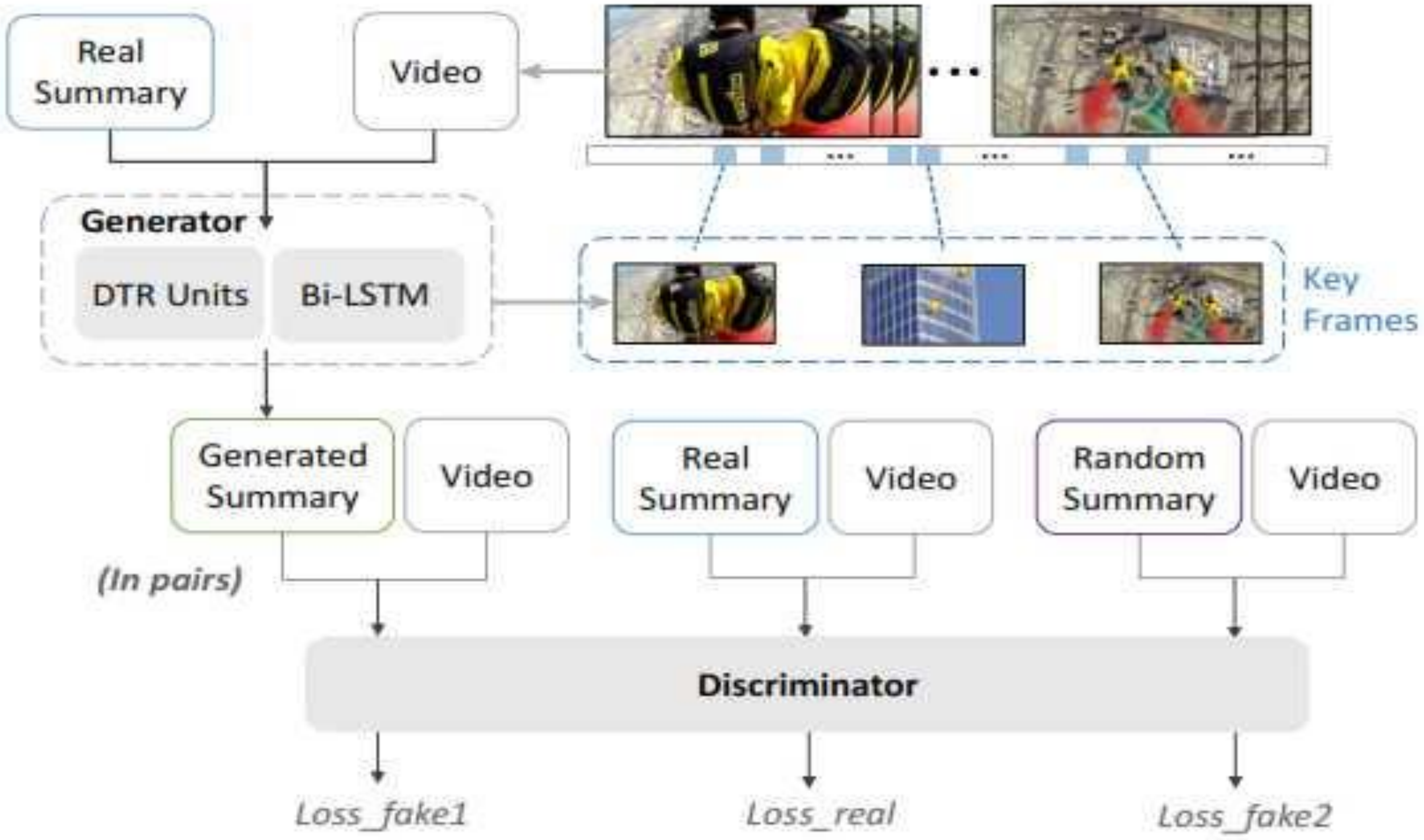
- Typically, $\sigma \in [0.1, 0.2]$.
- The sparsity loss updates $\mathbf{w}_s$.

DTR-GAN

The ***Dilated Temporal Relational Generative Adversarial Network*** (DTR-GAN) is an architecture slightly similar to SUM-GAN, but it is ***supervised***.

- The Discriminator in DTR-GAN is trained with a composite three-part loss function, that takes jointly into account the ***generated summary***, the ***ground-truth summary*** and a ***random summary***.
- This provides better regularization.

DTR-GAN



DTR-GAN (Image from [DIN2019])

Artificial Intelligence & Information Analysis Lab

DTR-GAN

- The Frame Selector is enhanced in DTR-GAN: besides the LSTMs, it also contains ***Dilated Temporal Relational*** (***DTR***) units.

- DTR units aim to exploit ***long-range temporal dependencies***, complementing LSTMs.

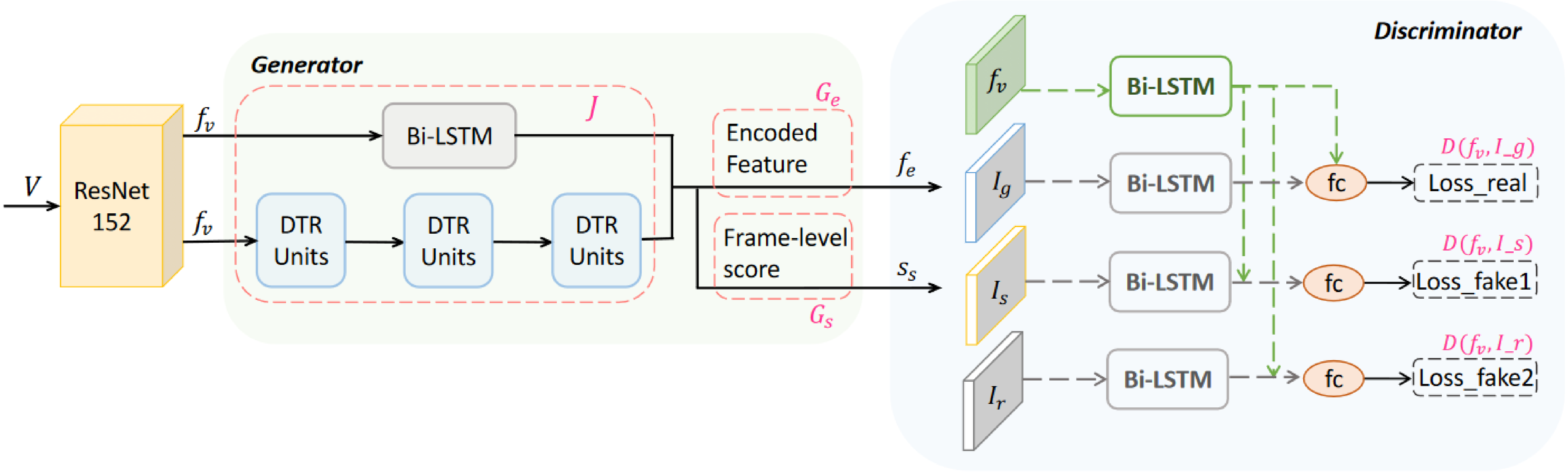
- They integrate context among video frames at multi-scale time spans, hence enlarging the temporal window and the temporal inter-frame relations.

DTR-GAN

- There is no LSTM auto-encoder in the DTR-GAN Summarizer, because the Discriminator is given ***video + summary pairs*** as inputs.

- The Discriminator learns to evaluate the correspondence between an input video and its summary, rather than input video reconstruction from its summary.

DTR-GAN



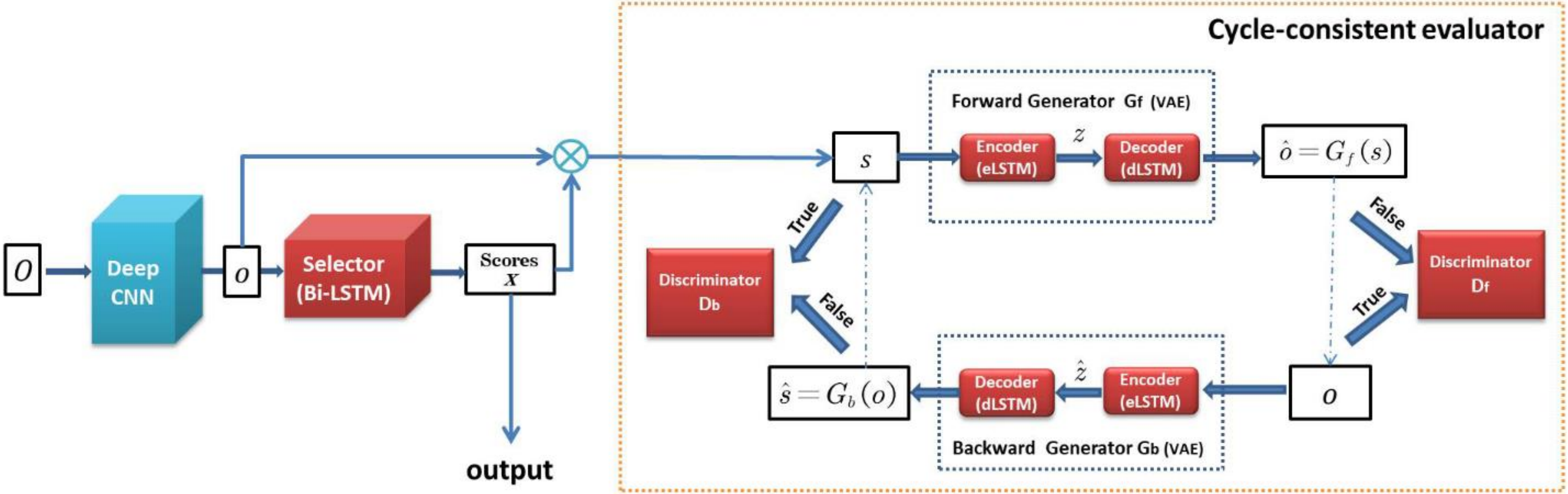
DTR-GAN (Image from [DIN2019])

Cycle-SUM

Cycle-SUM is an unsupervised end-to-end trainable DNN for key-frame extraction, which extends the original SUM-GAN.

- During training, it replaces the unidirectional reconstruction of SUM-GAN/SUM-GAN-AAE (the original video is reconstructed from the generated summary) with a ***"circular" bidirectional video reconstruction***.
- A ***cyclic consistency loss term*** is added to the training objectives of the overall framework.

Cycle-SUM



Cycle-SUM architecture. (Image from [PIN2019])

Cycle-SUM

- Cycle-SUM is composed of an initial Frame Selector, two autoencoders (instead of one) and two Discriminators (instead of one).

- The ***forward autoencoder*** and Discriminator ***reconstruct the original video*** from the generated summary and evaluate it, respectively.

- The ***backward autoencoder*** and Discriminator ***reconstruct the summary*** from the original video and evaluate it, respectively.

Cycle-SUM

- The closed training loop enforces the cyclic consistency.
- It aids the DNN to ***maximize mutual information*** between the summary and the original, full-length video.
- Explicitly enforcing the reconstruction cycle original → summary → original → summary, better guarantees summary completeness and representativeness.

Summary diversity

Emphasis of DNN-based video summarization methods:

- Summary representativeness, conciseness and completeness.
- However, it may be equally important that the selected key-frames are ***diverse in visual content***.
- Summary variety makes it summary more interesting and reduces redundancy.

Summary diversity

- A straightforward way to achieve ***summary diversity*** with DNNs is to add the so-called ***Determinantal Point Process*** (DPP) loss term in the pool of training objectives.

- In frameworks similar to SUM-GAN, the DPP loss directs the training process so that the Frame Selector learns to create a ***diverse overall summary***.

- This diversity pertains to the semantic content captured in the input video frame representations (e.g., visible objects).

Summary diversity

- The DPP loss operates by:
  - ***quantifying the variance*** of video frame representations.
  - penalizing candidate key-frame sets/summaries that do not capture significant percentage of the original video variance.
- Consider a matrix $\mathbf{L} \in \mathbb{R}^{T \times T}$ by computing the pairwise cosine similarity for video frames at time t and t':

$$L_{ij} = \mathbf{e}_t^T \mathbf{e}_{t'}.$$

- $\mathbf{e}_t, \mathbf{e}_{t'}$: Encoder hidden states at time t and t', respectively.

Summary diversity

DPP loss:

$$\mathcal{L}_{dpp} = -\log\left(\frac{\det(\mathbf{L}_y)}{\det(\mathbf{L}+\mathbf{I})}\right).$$

- $\mathbf{L}_y$ is a submatrix $\mathbf{L}$. Its rows and columns indicate the selected summary frames. $\mathbf{I}$ is the identity matrix.
- Recently, the DPP loss was extended to capture the ***diversity of additional modalities***, besides the CNN video frame representation,
- The diversity in the textual descriptions of each video frame, scene context and visible activities are enforced [KAS2022].

Summary diversity

- SUM-GAN-AAE is employed as a baseline and a pre-trained image captioner P is required.
- Then, the ***DPP-caption loss*** exhorts the video summary to be more diverse in terms of textual semantic content.
- During training, each video frame is forwarded to P, in parallel to feeding it to the Encoder.
- The following cost is used for Frame Selector weight update:

$$\mathcal{L}_{dpp-c} = -\log \frac{\det(\mathbf{P}_y)}{\det(\mathbf{P}+\mathbf{I})}.$$

Artificial Intelligence & Information Analysis Lab

DNNs and dictionary learning

Dictionary learning in unsupervised DNN frameworks, such as SUM-GAN-AAE, has also been attempted [KAS2021].

- Using SUM-GAN-AAE as a baseline, an additional ***pre-trained autoencoder*** encodes the entire video sequence into vector $\mathbf{h}$.

- During training, a novel loss term is added to the framework:

$$\mathcal{L}_{dict} = \|\mathbf{h} - \mathbf{A}\mathbf{e}\|_2.$$

- $\mathbf{A}$ essentially serves as a ***global visual dictionary***.

- Vector $\mathbf{e}$ is given by the Encoder, while $\mathbf{A}$ is learnt.

DNNs and dictionary learning

- Matrix $\mathbf{A}$ transforms the current summary representation to a vector space being simultaneously learnt from all the original videos.

- Thus, each summary representation is exhorted towards being a set of linear reconstruction coefficients that are jointly able to reproduce the corresponding original video representation.

- This is on top of the non-linear reconstruction objective enforced by the baseline SUM-GAN-AAE.

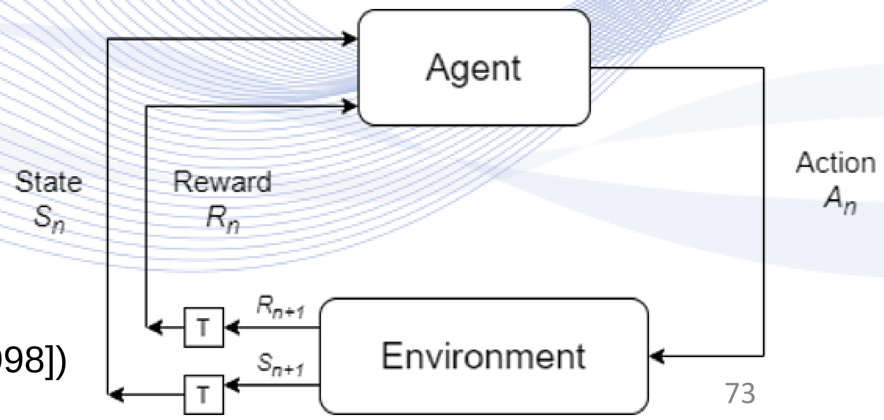
DNNs and reinforcement learning

- ***Reinforcement learning*** (RL) has also been integrated into unsupervised deep neural frameworks for video summarization.

- In RL, a cognitive agent is trained through interaction: it interacts with its environment, in order to find a policy that maximizes a ***cumulative reward***.

- The reward is a numerical measure that determines how good the agent action was.

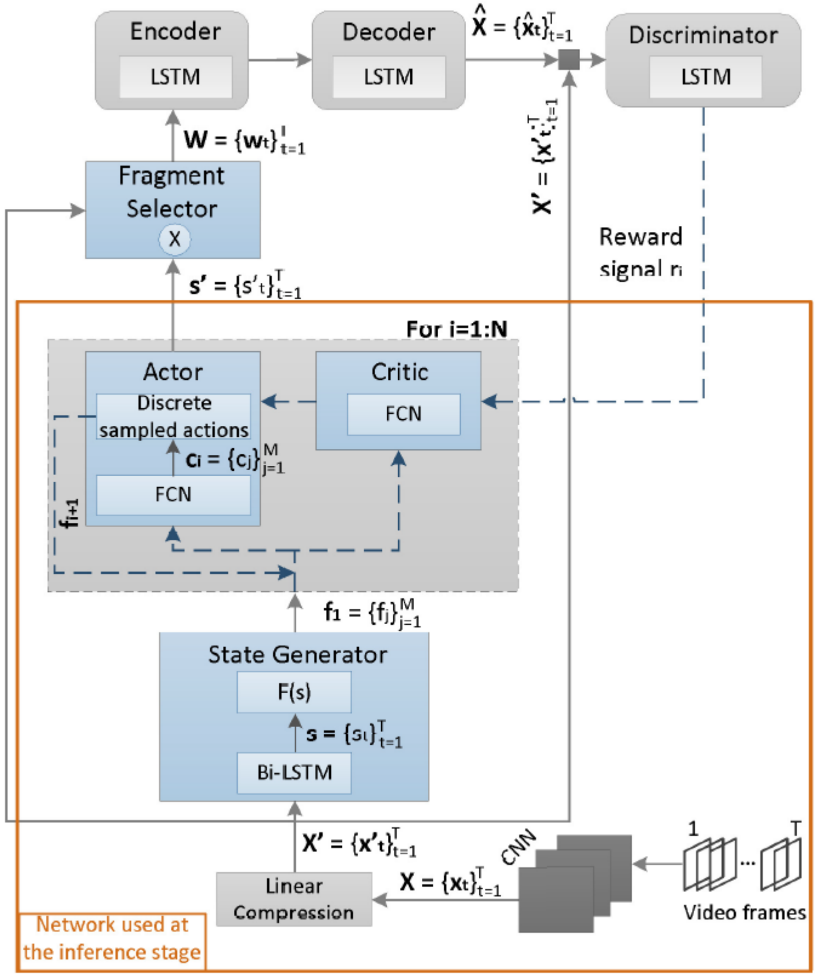
- The learned policy maps states to actions.



Environment-action interaction (Image from [SUT1998])

DNNs and reinforcement learning

- AC-SUM-GAN is a good example of combining SUM-GAN with RL [APO2020].
- A ***neural Actor-Critic architecture*** is embedded into SUM-GAN.
- During training, it learns the optimal policy for key-frame extraction.
- During inference, the RL agent modifies/adjusts the video frame importance scores outputted by the Frame Selector.



The architecture of AC-SUM GAN (Image from [APO2020])

DNNs and reinforcement learning

- ***The Actor generates sequences incrementally***, based on a set of discrete sampled actions over a group of video fragments.

- ***The Critic evaluates the Actor choices*** and returns a value for scoring each choice, according to its impact on the action-state space.

- The ***Discriminator*** acts as the RL environment and ***returns a reward that is used to train the Actor-Critic model***, which learns a value function (Critic) and a policy for key-fragment selection (Actor).

- The Critic can be discarded after training.

DNNs and reinforcement learning

- The Actor plays an ***N-picks game*** to explore the action-state space.

- For every step i, $(1 \leq i \leq N)$:

  - It receives the current state $\mathbf{f}_i = \{f_j\}_{j=1}^{M}$, where M is the number of non-overlapping fragments into which the video is segmented.
    - At time $i = 1$, $\mathbf{f}_1$ is derived from the vector of importance scores ***outputted by the Frame Selector***.

DNNs and reinforcement learning

- (continued)
  - It produces a ***distribution of actions*** $\mathbf{c}_i = \{c_j\}_{j=1}^{M}$.
  - It takes an action by sampling the computed distribution $\mathbf{c}_i$, thus, picking a video fragment k for inclusion in the summary.
  - This action modifies the state and produces $\mathbf{f}_{i+1}$.
  - During training, the reward is the Discriminator's classification decision.

Transformer Video Summarization

- Few works exist that use Transformers for Video Summarization.

- The input of the Transformer is a sequence of encoded video frames.
- The output of the Transformer network, is an ***importance score*** for each input frame or for short frame sequences.

- By thresholding the score, we can produce a video skim or multiple key-frames from the input video.

Transformer Video Summarization

- A 2D CNN (e.g., a pre-trained VGG [VGG2015]) is typically used to convert the input video into a sequence of vectors (one per video frame).

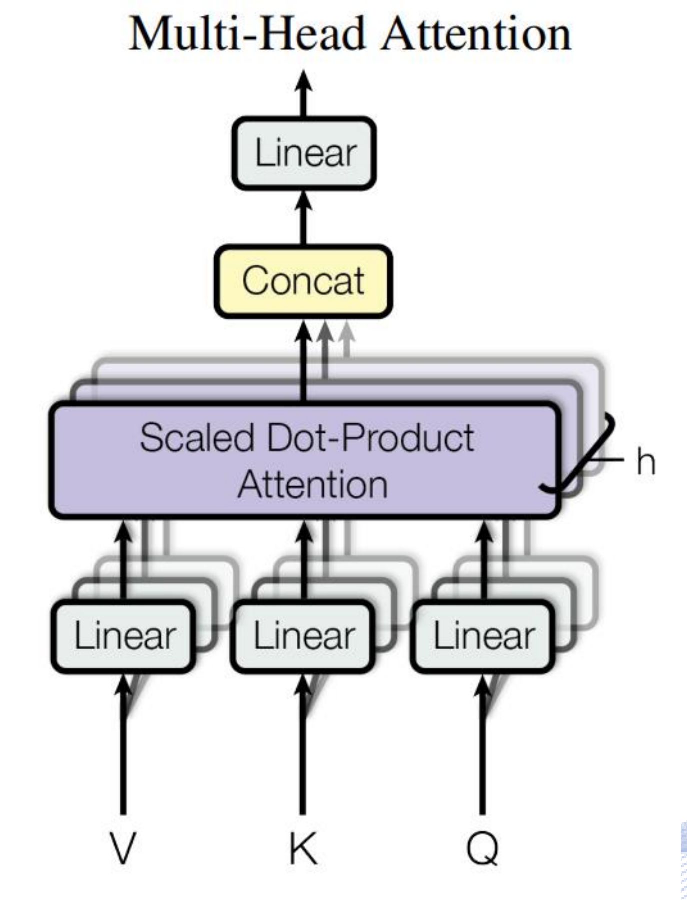
- Since the input is in the form of a sequence of vectors, well-known Transformers utilized in Natural Language Processing (e.g., BERT [BERT2018]) can be used to update the video frame representations.

Transformer Video Summarization

- The basic building component of the transformer is the ***Self-Attention*** (SA) block.
  - It is applied on entire video frames.
- Each SA block updates a vector representation, by implementing a weighted sum of all the sequence representations:
  - Similar vectors produce a summation weight (***attention coefficient***) close to 1.
  - Dissimilar vectors produce a weight close to 0.

Transformer Video Summarization

- Attention operation:
  - $A(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax\left(\frac{\mathbf{QK}^T}{\sqrt{d_k}}\right)\mathbf{V}$
    - $\mathbf{Q} = \mathbf{W}_Q\mathbf{X}$
    - $\mathbf{K} = \mathbf{W}_K\mathbf{X}$
    - $\mathbf{V} = \mathbf{W}_V\mathbf{X}$.
  - $\mathbf{X}$: input sequence
  - $\mathbf{W}_Q, \mathbf{W}_V, \mathbf{W}_K$: learnable weight matrices.



Multi-Head Attention block [BERT2018].

Artificial Intelligence & Information Analysis Lab

Transformer Video Summarization

- Pros:
  - The Transformer is permutation equivariant. It can model relationships between sequence tokens (i.e., encoded video frames) even if the input sequence is arranged in a different way.
  - Since every sequence token “attends” to all other tokens the same way, the Transformer can discover relationships between parts of the sequence, regardless of their proximity.

Transformer Video Summarization

- Cons:
  - The Transformer has $O(N^3)$ computational complexity, due to the SA operator.
  - The most common SA implementation, consists of multiple heads with every head having a separate set of weights $(\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V)$. Since each block usually has multiple heads and the architecture consists of sequential blocks, the number of the network parameters can get very high. High complexity often leads to slow training times and produces a tendency to overfitting.

Multi-Modal Video Summarization

- A generic video summary is an abridged version of a video that conveys the whole story and features the most important scenes.

- Yet, the importance of video scenes is often subjective, and users should have the option of customizing the summary by using natural language to specify what is important to them.

- The seminal paper that utilized transformers to combine natural language with visual information for Video Summarization is “CLIP-It! Language-Guided Video Summarization” [CLIP2021].

Multi-Modal Video Summarization

[CLIP2021] Given a day-long video of a national park tour, the generic summary (top) is a video with relevant and diverse keyframes. When using the query “All the scenes containing restaurants and shopping centers”, the generated query-focused summary includes all the matching scenes. Similarly, the query “All water bodies such as lakes, rivers, and waterfalls”, yields a short summary containing all the water bodies present in the video.

Multi-Modal Video Summarization

Clip-it! Image - text combination methodology overview:

- Given a video of N frames, feature embeddings for each frame are extracted by a pre-trained DNN f_{img}.

- If a query is provided (in the form of a natural language string), it is embedded using a pretrained network f_{txt}.

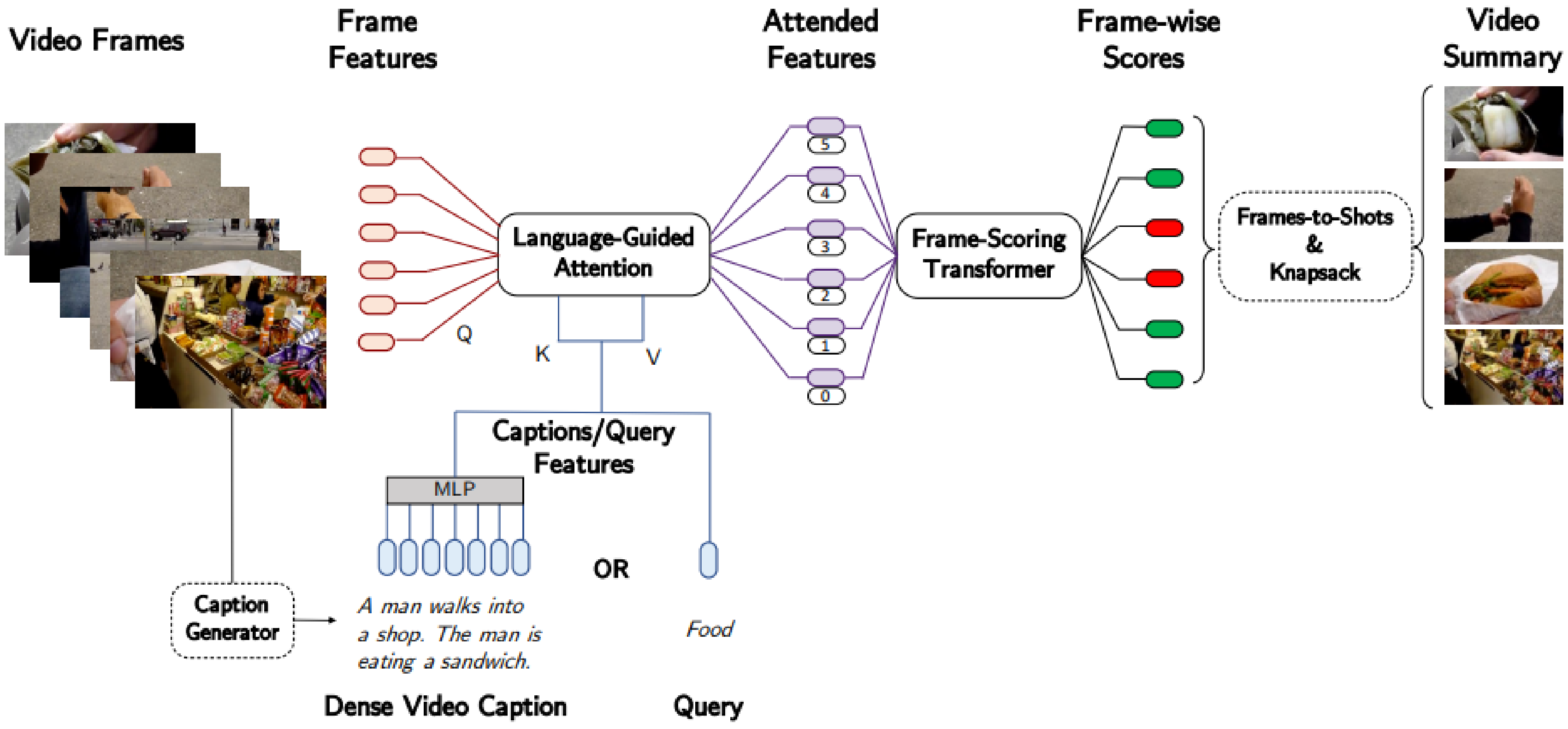
- Alternatively, an off-the-shelf video captioning model is used to generate a dense video caption with M sentences, where $M = N$. Then, the sentences are embedded using f_{txt}.

Multi-Modal Video Summarization

- Next, language attended image embeddings I^* are computed, using learned Language-Guided Multi-head Attention f^*_{imgtxt}.

- Finally, Frame-Scoring Transformer is trained which assigns scores to each frame in the video

Multi-Modal Video Summarization



Clip-it! Transformer [CLIP2021].

Evaluation Datasets

- There are several public datasets for evaluating video summarization algorithms.
- Typically, these datasets provide a collection of videos with associated per-frame ground truth importance scores.
- The most common ones are TVSum and SumMe.
  - ***SumMe*** includes 25 videos of 1 to 6 minutes duration with diverse video contents, captured both from first and third-person view.
  - ***TVSum*** consists of 50 videos of 1 to 11 minutes duration, containing video content from 10 categories of the TRECVid MED dataset.

Evaluation Datasets

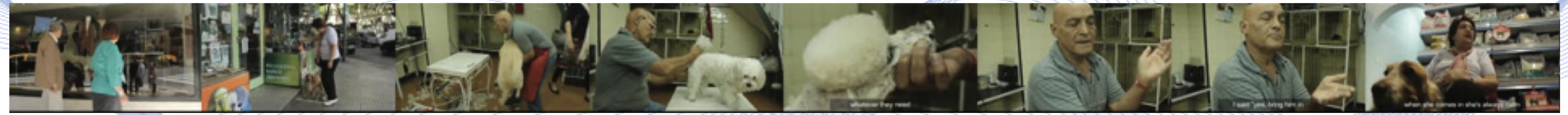
- Every video of the dataset ***is annotated by multiple users*** in the form of key fragments (SumMe) or frame-level importance scores (TVSum)
  - Single ground-truth summaries are also provided.
- To evaluate a video summarization algorithm, the generated summary for a given video is compared with the users' summary, separately per user.

Evaluation Datasets

- An ***F-Score*** (F-measure) is computed for each pair of compared summaries.
- The computed F-Scores for TVSum are averaged or the maximum of them is kept for SumMe and a final F-Score is obtained for this video.
- The computed F-Scores for the entire set of testing videos are finally averaged to quantify the algorithm’s performance.

Evaluation Datasets

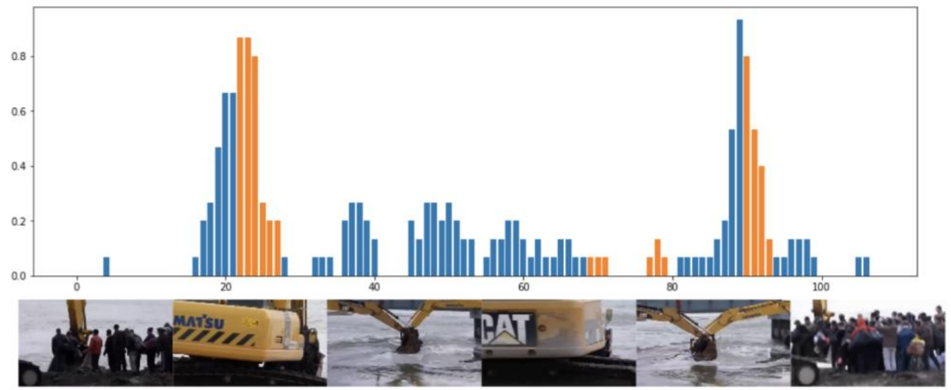
Video frames from the sequence “Cooking” of the SumMe dataset.



Video frames from the sequence “Dog grooming in Buenos Aires” of the TVSum dataset.

Evaluation Datasets

Video frames from the sequence “Excavators road crossing” of the SumMe dataset.



Video frame importance scores and the extracted summary using SUM-GAN-AAE in combination with $\mathcal{L}_{dict} + \mathcal{L}_{dpp}$.

Bibliography

[PIT2017] I. Pitas, “Digital video processing and analysis” , China Machine Press, 2017 (in Chinese).
[PIT2013] I. Pitas, “Digital Video and Television” , Createspace/Amazon, 2013.
[PIT2021] I. Pitas, “Computer vision”, Createspace/Amazon, in press.
[NIK2000] N. Nikolaidis and I. Pitas, “3D Image Processing Algorithms”, J. Wiley, 2000.
[PIT2000] I. Pitas, “Digital Image Processing Algorithms and Applications”, J. Wiley, 2000.

Bibliography

[DAR2014] K. Darabi and G. Ghinea, “Personalized video summarization by highest quality frame”, IEEE International Conference on Multimedia and Expo Workshops (ICMEW), 2014.

[ZHA2006] Z. Zhao, S. Jiang, Q. Huang and G. Zhu, “Highlight summarization in sports video based on replay detection”, IEEE International Conference on Multimedia and Expo, 2006.

[BOR2018] A. Bora and S. Sharma, “A review on video summarization approaches: Recent advances and directions”, International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018.

[IRI2010] G. Irie, T. Satou, A. Kojima, T. Yamasaki and K. Aizawa, ”Automatic trailer generation”, ACM International Conference on Multimedia, 2010.

[KAS2022] M. Kaseris, I. Mademlis, and I. Pitas, “Exploiting caption diversity for unsupervised video summarization“, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[MAD2018] I. Mademlis, A. Tefas, and I. Pitas, “A salient dictionary learning framework for activity video summarization via key-frame extraction”, Elsevier Information Sciences, 432, 319-331, 2018.

Bibliography

[BUR2020] H. B.U. Haq, M. Asif and M. B. Ahmad, “Video summarization techniques: A review”, International Journal of Scientific Technology Research”, volume 9, issue 11, 2020.

[KAI2012] G. Guan, Z. Wang, K. Yu, S. Mei, M. He and D. Feng, “Video summarization with global and local features”, IEEE International Conference on Multimedia and Expo Workshops, 2012.

[SAB2012] W. Sabbar, A. Chergui, A. Bekkhoucha, “Video summarization using shot segmentation and local motion estimation”, Innovative Computing Technology, pp. 190–193, 2012.

[MAD2016] I. Mademlis, A. Tefas, N. Nikolaidis and Ioannis Pitas, “Multimodal stereoscopic movie summarization conforming to narrative characteristics”, IEEE Transactions on Image Processing, 25.12: 5828-5840, 2016.

[SUT1998] R. S. Sutton and A. G. Barto, “An Introduction to Reinforcement Learning”, MIT Press, 1998.

Bibliography

[WOR2020] A. Workie and R. Sharma, “Digital video summarization techniques: A survey”, International Journal of Engineering Research Technology, vol. 9, issue 01, pp. 81-85, 2020.

[SUP2017] J. Supancic III, D. Ramanan. “Tracking as online decision-making: Learning a policy from streaming videos with reinforcement learning”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[TRU2007] B. T. Truong, Venkatesh s., “Video Abstraction: A Systematic Review and Classification”, ACM Transactions on Multimedia Computing, Communications, and Applications, 3:3, 2007.

[XIA2021] H. Xiao and J. Shi, “Diverse video captioning through latent variable expansion”, arXiv preprint arXiv:1910, 2021.

[KAS2021] M. Kaseris, I. Mademlis and I. Pitas, "Adversarial Unsupervised Video Summarization Augmented With Dictionary Loss“, Proceedings of the IEEE International Conference on Image Processing (ICIP), 2021.

Bibliography

[SHE2015] C. V. Sheena, N. K. Narayanan, “Key-frame extraction by analysis of histograms of video frames using statistical methods”, International Conference on Eco-friendly Computing and Communication Systems, 2015.

[BAL2019] D. Sen and B. Raman, “Video skimming: Taxonomy and comprehensive survey”, arXiv preprint arXiv:1909.12948, 2019.

[METS2020] I. A. Metsai, V. Mezaris, E. Apostolidis, E. Adamantidou and I. Patras, “Unsupervised video summarization via attention-driven adversarial learning”, International Conference on Multimedia (MMM), 2020.

[DIN2019] D. Zhang, M. Tan, E. P. Xing ,Y. Zhang, X. Liang, “Dilated temporal relational adversarial network for generic video summarization”, Springer Multimedia Tools and Applications, 2019, 78.24, pp. 35237-35261.

[MAH2017] B. Mahasseni, M. Lam and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks“, Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017.

Bibliography

[PIN2019] P. Li, L. Zhou, J. Feng, L. Yuan, F. EH Tay, “Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization”, Proceedings of the AAAI Conference on Artificial Intelligence, 2019.

[APO2020] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, I. Patras, “AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization”, IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31.8: 3278-3292.

[VGG2015] Simonyan, Karen and Zisserman, Andrew. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014)

[BERT2018] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton and Toutanova, Kristina, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” (2018). arxiv:1810.04805

[CLIP2021] M. Ranzato and A. Beygelzimer and Y. Dauphin and P.S. Liang and J. Wortman Vaughan, “CLIP-It! Language-Guided Video Summarization”, Advances in Neural Information Processing Systems, 2021, 34: 13988 -14000.

Q & A

Thank you very much for your attention!

More material in
http://icarus.csd.auth.gr/cvml-web-lecture-series/

Contact: Prof. I. Pitas
pitas@csd.auth.gr