Selecting a Diverse Set of Aesthetically-Pleasing and Representative Video Thumbnails Using Reinforcement Learning

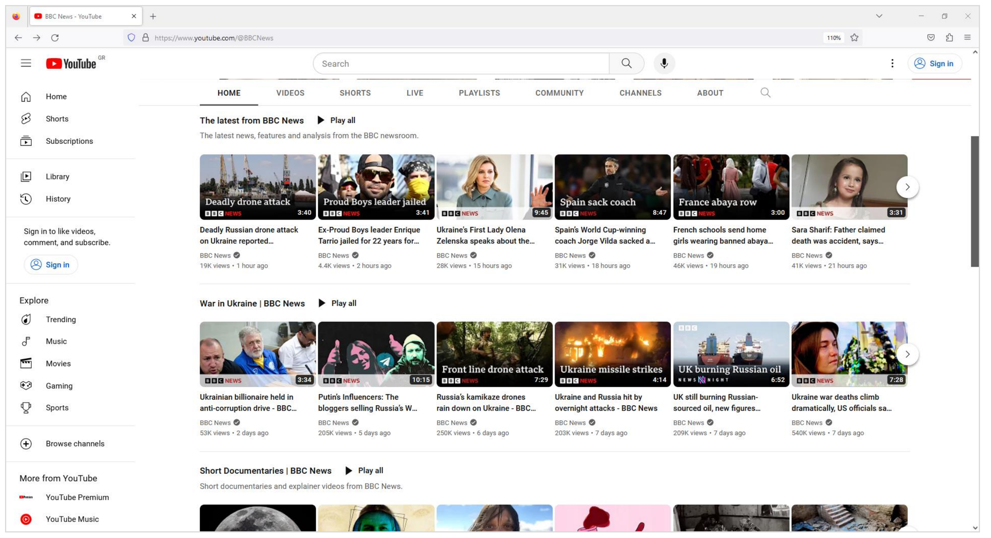
Evlampios Apostolidis[1,2], Georgios Balaouras[1], Vasileios Mezaris[1], Ioannis Patras[2]

[1] Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece
[2] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK
{apostolid, mpalaourg, bmezaris}@iti.gr, i.patras@qmul@ac.uk

Video thumbnail selection

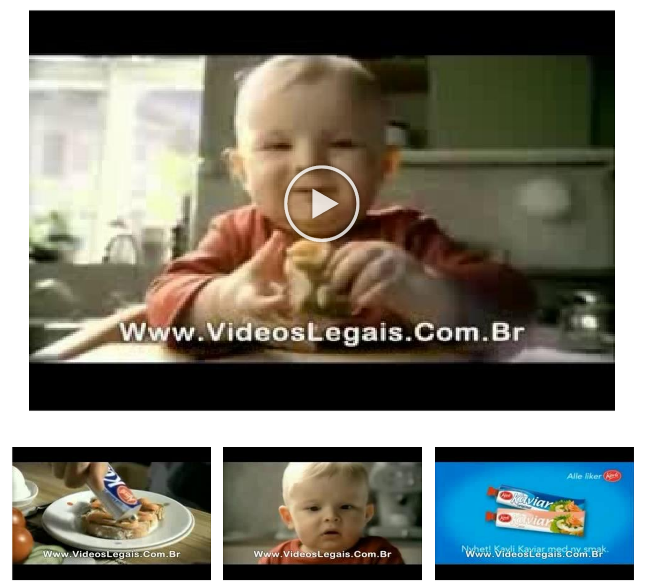
Motivation

- **Tremendous growth of videos** over the Web
- How to easily **find what we are looking for**?
- Video sharing platforms & social networks **represent videos using thumbnails**
- Manual thumbnail selection is a **tedious & time-consuming process**



Goal

"Given a video, select one or a few video frames that provide a ***representative & aesthetically-pleasing overview*** *of its content"*



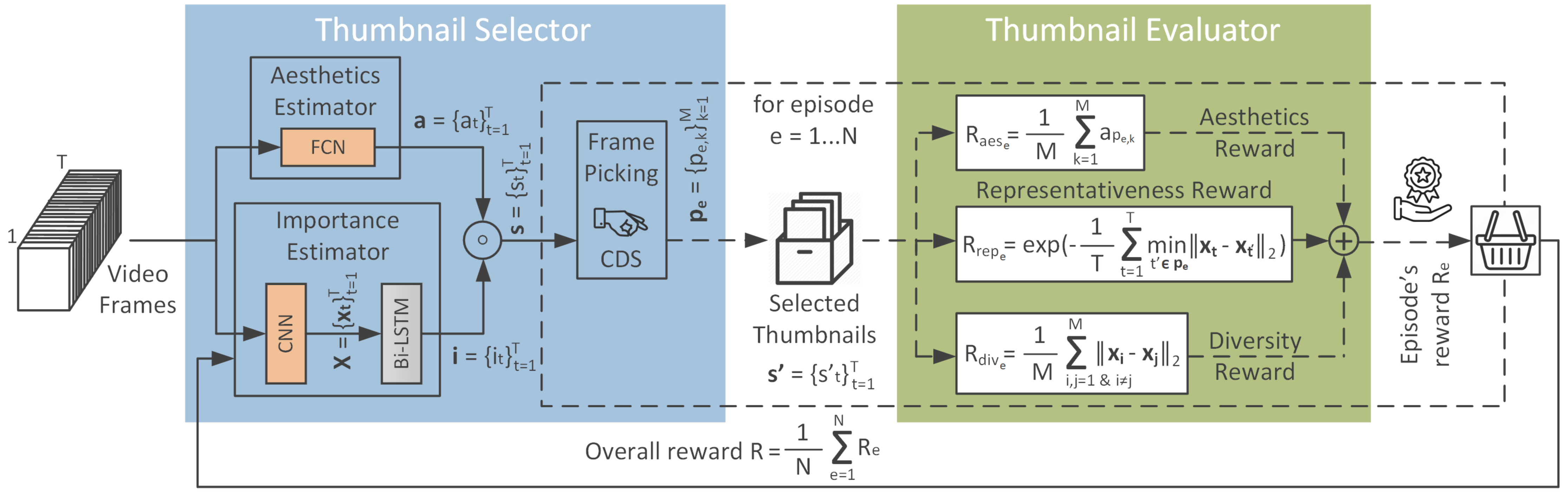
Existing (visual-based) solutions

- **Early approaches:** use rules about the thumbnail & extract low-level (luminance) & mid-level features (faces) to assess frames' alignment with them
- **Recent approaches:** focus on the aesthetic quality & representativeness of frames, & are based on: i) feature extraction & clustering, or ii) deep networks

Proposed method: RL-DiVTS

Thumbnail Selector (used during training & inference)

- **Aesthetic Estimator**: scores frames based on their aesthetic quality (pretrained FCN on AVA)
- **Importance Estimator**: scores frames by modeling their temporal dependence (pretrained CNN on ImageNet & trainable bi-directional LSTM)
- **Frame Picking Mechanism**: picks frames sequentially by sampling from a categorical distribution & demoting the selection of frames similar to the already picked ones



Training approach: Episodic REINFORCE algorithm

Thumbnail Evaluator (used only during training)

- Assesses the selected thumbnails in terms of **aesthetic quality**, **representativeness** & **diversity**, using three tailored reward functions
- The **overall reward per episode** is formed by: $R_e = a \cdot R_{aes_e} + \beta \cdot D \cdot R_{rep_e} + \gamma \cdot R_{div_e}$ (D projects R_{rep_e} in the same scale with the other rewards)
- The **average reward across all episodes** formulates its feedback for the current training sample

Experimental results

Experimental setting

- **Datasets:** OVP (50 videos) & YouTube (50 videos)
- **Data spit:** 80% training & 20% testing
- **Ground-truth:** 3 most selected keyframes by humans
- **Evaluation approach:** "top-3 matching" (overlap between ground-truth & selected thumbnails)
- **Similarity** with ground-truth thumbnails: measured by SSIM (declare a "match" if SSIM > 0.7)

Comparison of RL-DiVTS with other approaches

- **Performs consistently well** on both datasets (best & second best-performing one)
- Is **more effective** compared to methods for **video summarization** (AC-SUM-GAN, CA-SUM)
- Is **significantly better than ARL-VTS** (our previous method) in terms of **performance**, **training time** & **memory footprint**

	OVP	YouTube
Baseline (Random)	8.63 ± 2.50	4.41 ± 1.77
AC-SUM-GAN	7.87 ± 3.41	7.33 ± 0.70
CA-SUM	7.60 ± 2.85	8.00 ± 3.56
Hecate-VTS	11.72	16.47
ReconstSum	12.18	**18.25**
ARL-VTS	12.50 ± 3.37	7.83 ± 1.49
RL-DiVTS (proposed)	**25.33** ± 3.97	17.50 ± 2.57

	Training time (sec/epoch)		# Param.
	OVP	YouTube	(in Millions)
ARL-VTS	38.41	62.43	28.36
RL-DiVTS	2.33	2.70	12.60

Ablation study

- Removal of either of the **used criteria** & the **Frame Picking mechanism** leads to **reduced performance** in, at least, one of the datasets

	OVP	YouTube
RL-DiVTS w/o AES	14.13 ± 2.96	10.33 ± 1.73
RL-DiVTS w/o REP	20.53 ± 1.91	13.17 ± 1.09
RL-DiVTS w/o DIV	**26.40 ± 1.30**	14.33 ± 1.49
RL-DiVTS w/o CDS	24.67 ± 3.16	15.00 ± 1.44
RL-DiVTS (proposed)	25.33 ± 3.97	**17.50 ± 2.57**

Software available at: https://github.com/e-apostolidis/RL-DiVTS

This work was supported by the EU Horizon 2020 programme under grant agreement H2020-951911 AI4Media