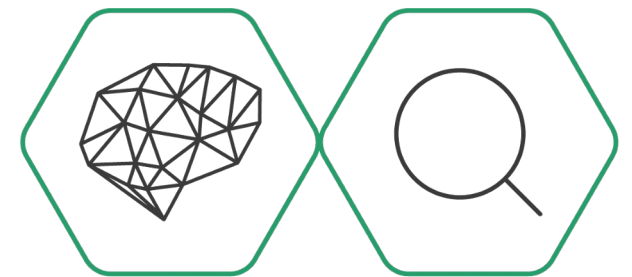


Keyframe Selection and Enhancement service

Kostas Apostolidis, Vasileios Mezaris

CERTH



vera.ai

Overview



- The problem
- Our approach
- How it works
- Key technologies
- Workflow
- The KSE service

The problem





Is that video
fake?



“Would the famous Eiffel Tower in
#Paris ... remain standing under
endless bombing of Russian
troops? ... Today it's #Ukraine,
tomorrow it will be the whole of
#Europe.”



Is this taken
out-of-context?



Media Verification



Synthetic videos are not the only concern!

- Content Alteration: Were original videos used as a basis?
- Contextual Misuse: Were videos presented out of context?

Media Verification



Synthetic videos are not the only concern!

- Content Alteration: Were original videos used as a basis?
- Contextual Misuse: Were videos presented out of context?

Media Verification



Synthetic videos are not the only concern!

- Content Alteration: Were original videos used as a basis?
- Contextual Misuse: Were videos presented out of context?

A real video that has been digitally altered to change its content, often by adding, removing, or modifying visual/audio elements.

For example, a politician's speech is edited to change their words, making it appear as if they said something controversial.

Media Verification



Synthetic videos are not the only concern!

- Content Alteration: Were original videos used as a basis?
- Contextual Misuse: Were videos presented out of context?

Media Verification



Synthetic videos are not the only concern!

- Content Alteration: Were original videos used as a basis?
- Contextual Misuse: Were videos presented out of context?

A real video that is presented with misleading captions, timestamps, or contextual framing to support a false narrative.
For example, a video of military exercises is falsely claimed to show an ongoing invasion, causing unnecessary panic.



Our approach

AI-powered keyframe selection and enhancement



- Find critical frames in video
- Reduce video to keyframes
 - Serve as the initial stage in most video analysis pipelines
 - Enable subsequent analysis of video from other tools, e.g. reverse image search)
- Detect key elements for closer inspection
- Turn pixelated text/faces into readable details/clearer faces

AI-powered keyframe selection and enhancement



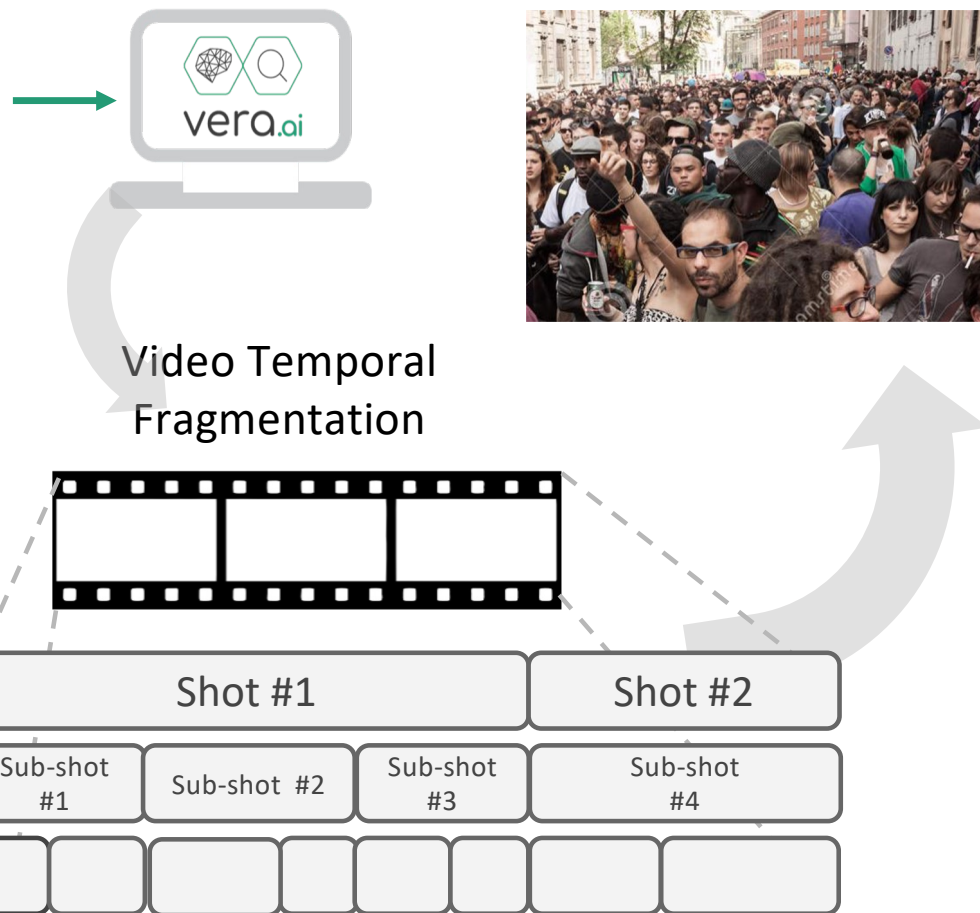
- Not a perfect solution, more of a “force multiplier”



How it works

AI-powered keyframe selection and enhancement

Process Overview



- **Fragment** video to shot and subshots, selecting **representative** keyframes for each fragment

AI-powered keyframe selection and enhancement

Process Overview



- Highlights parts of videos where useful clues might appear by **detecting** faces and text regions

AI-powered keyframe selection and enhancement

Process Overview





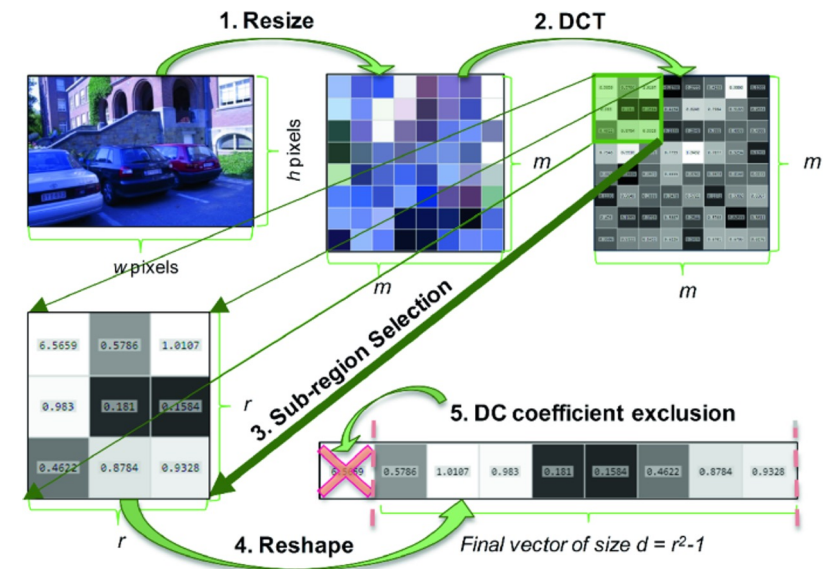
Key technologies

AI-powered keyframe selection and enhancement

Video fragmentation



- Video fragmentation is the process of dividing a video into meaningful segments, by detecting visual content changes, camera cuts, or significant content variations

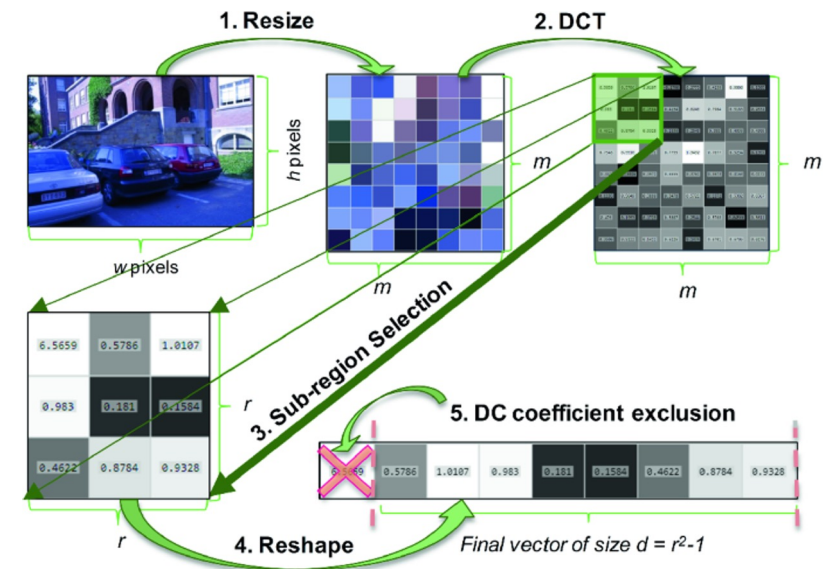


AI-powered keyframe selection and enhancement

Video fragmentation



- Video fragmentation helps break down videos into analyzable parts:
 - Scenes
 - **Shots**
 - Subshots

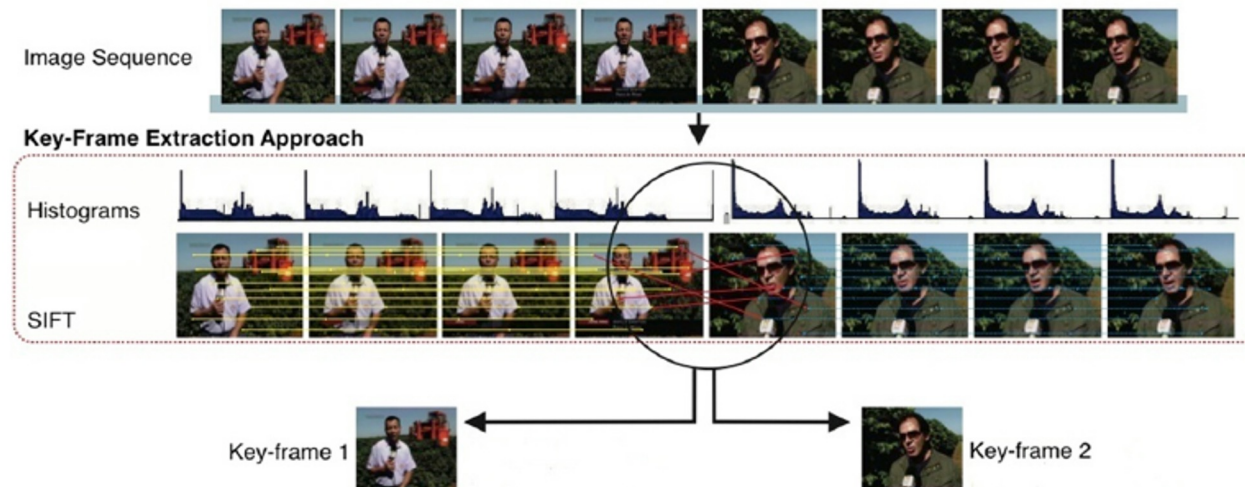


AI-powered keyframe selection and enhancement

Video Keyframe Selection



- Video Keyframe Selection refers to the process of identifying and extracting the most representative frames from a video to summarize its content while minimizing redundancy

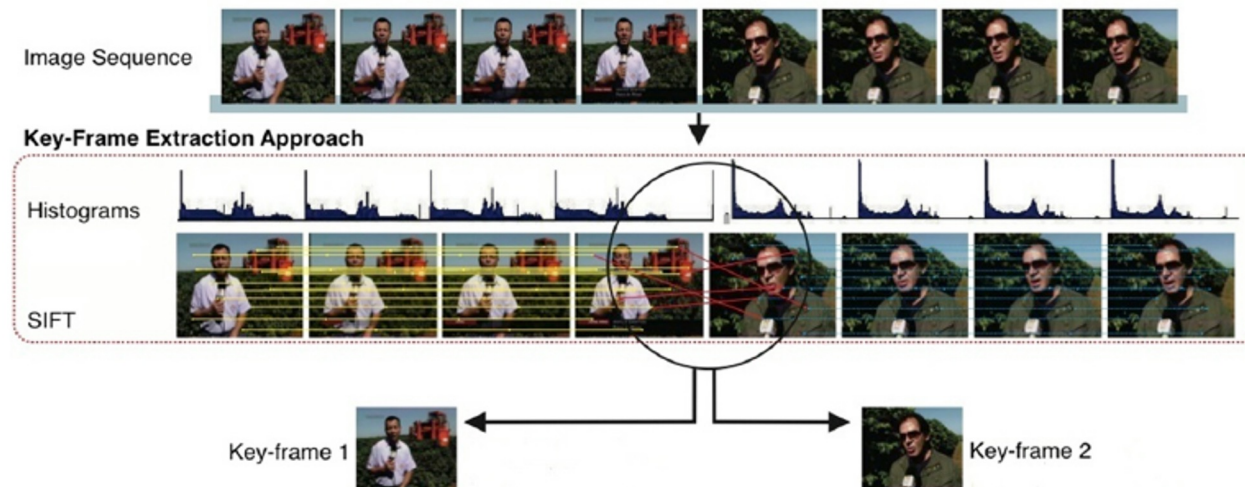


AI-powered keyframe selection and enhancement

Video Keyframe Selection



- User feedback: “The more, the merrier!”

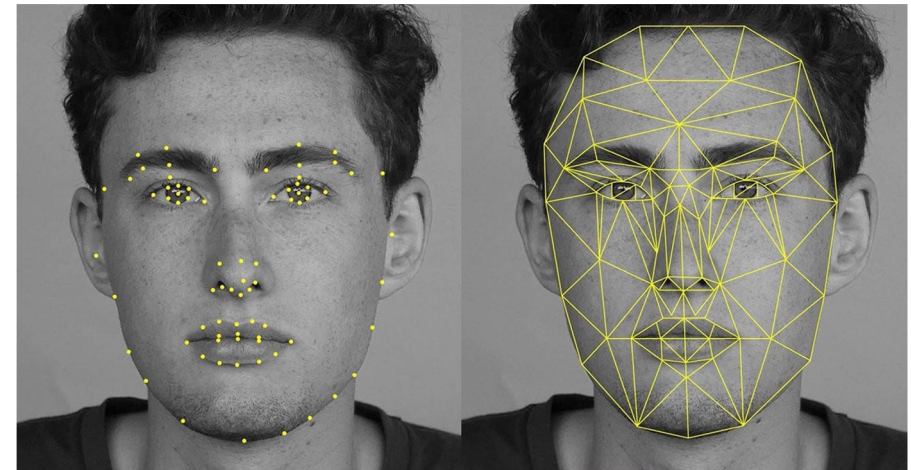


AI-powered keyframe selection and enhancement

Face detection



- Face detection in videos involves detecting and localizing human faces across video frames
- Detected faces can be compared across different videos to determine if the same footage has been repurposed

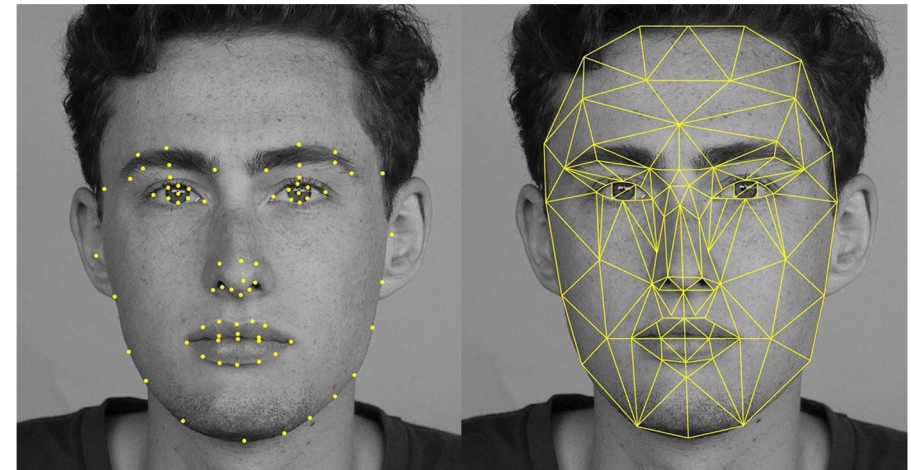


AI-powered keyframe selection and enhancement

Face detection



- Recent advancements in machine learning have enabled face detection algorithms to **surpass human performance** in accuracy and speed, particularly in challenging conditions such as low lighting, occlusions, etc.



AI-powered keyframe selection and enhancement

Text detection



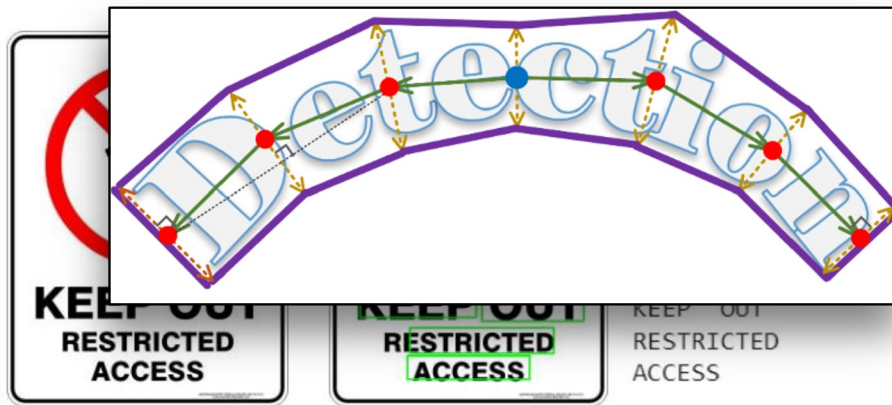
- Text detection in videos involves identifying and localizing text regions or video frames
- Extracted text can be compared across different videos to check if the same information appears elsewhere, indicating possible content repurposing

AI-powered keyframe selection and enhancement

Text detection



- Text detection has recently improved, especially for curved and irregular text



AI-powered keyframe selection and enhancement

Super-resolution



- Super-resolution is the process of enhancing the resolution of an image or video frame, typically using deep learning to reconstruct finer details from low-resolution inputs



Low-resolution image



High-resolution image

AI-powered keyframe selection and enhancement

Super-resolution



- Some super-resolution methods are specialized for text and faces, focusing on details that general approaches may overlook



Low-resolution image



High-resolution image

AI-powered keyframe selection and enhancement

Super-resolution



- Some super-resolution methods are specialized for text and faces, focusing on details that general approaches may overlook



AI-powered keyframe selection and enhancement

Super-resolution



- Some super-resolution methods are specialized for text and faces, focusing on details that general approaches may overlook

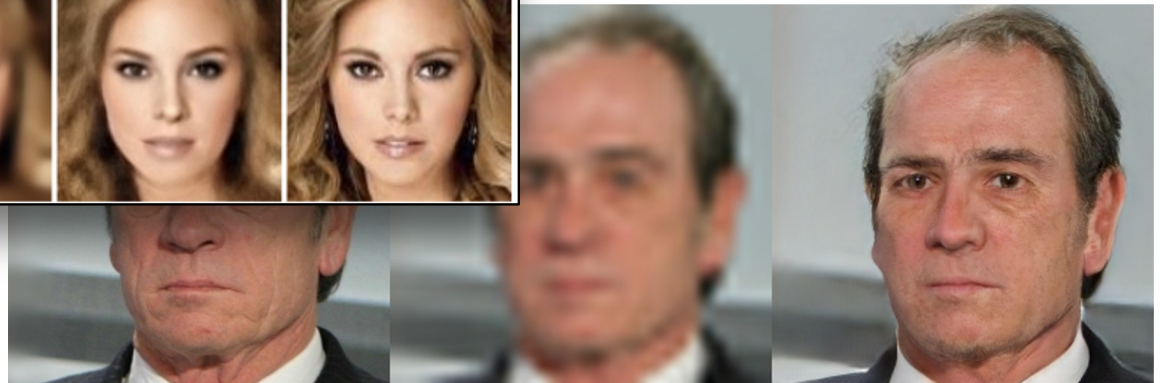
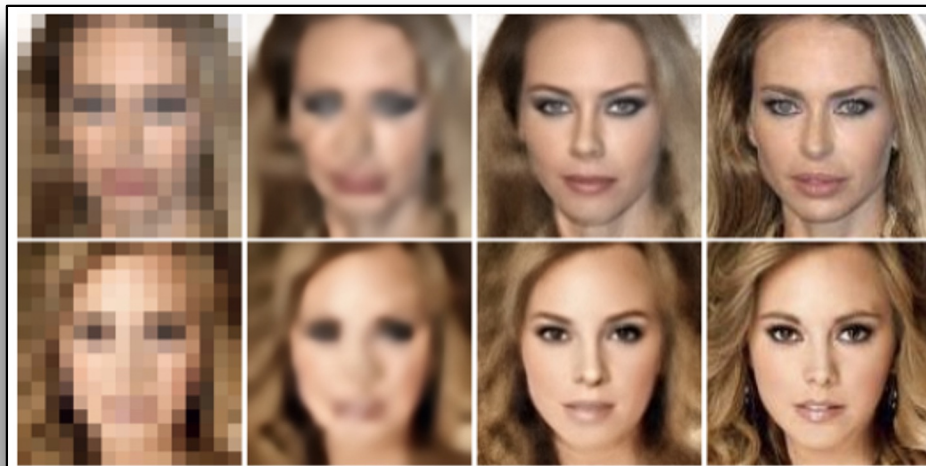


AI-powered keyframe selection and enhancement

Super-resolution



- Face super-resolution can create artificial details, known as hallucinations, that may alter facial features





The workflow

AI-powered keyframe selection and enhancement Workflow

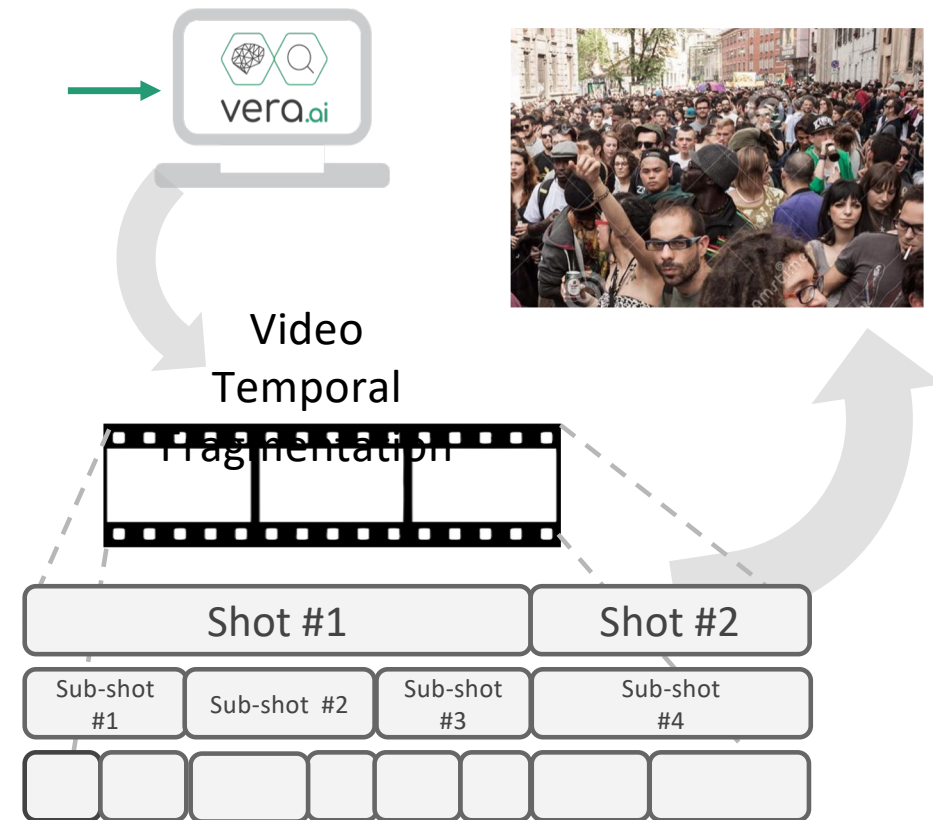


- Group visually very similar frames and process the least blurry from each group
- Fragment video to shots and subshots
- Detect key elements (faces, text)
- Group similar key elements by tracking the motion of the detections' bounding box
- Estimate the blurriness of each key element in the group and select the sharpest item
- Employ super-resolution techniques to enhance the selected item

AI-powered keyframe selection and enhancement Workflow



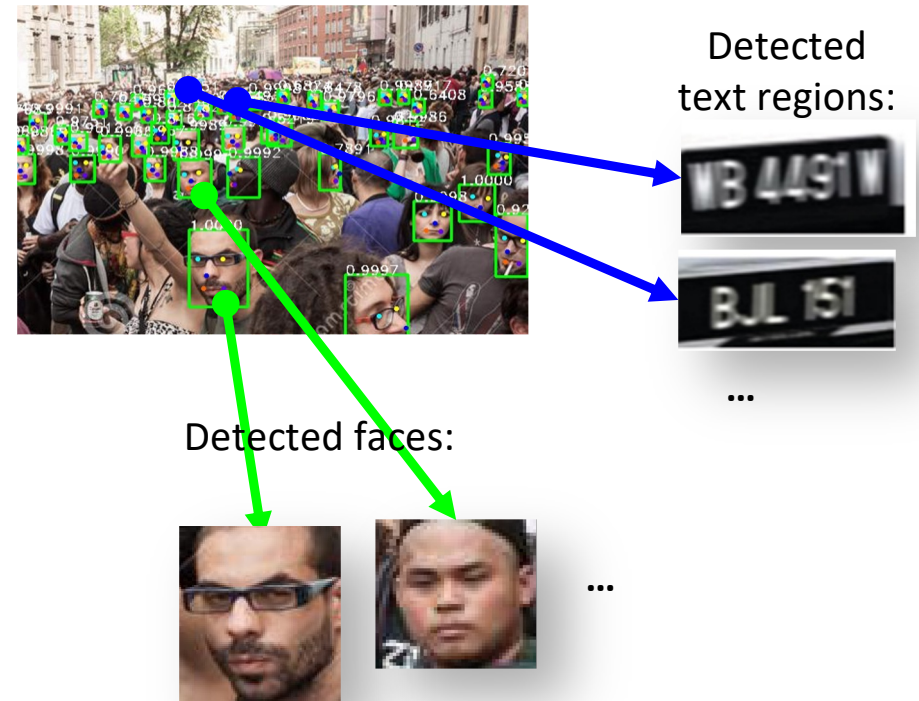
- Group visually very similar frames and process the least blurry from each group
- **Fragment video to shots and subshots**
- Detect key elements (faces, text)
- Group similar key elements by tracking the motion of the detections' bounding box
- Estimate the blurriness of each key element in the group and select the sharpest item
- Employ super-resolution techniques to enhance the selected item



AI-powered keyframe selection and enhancement Workflow



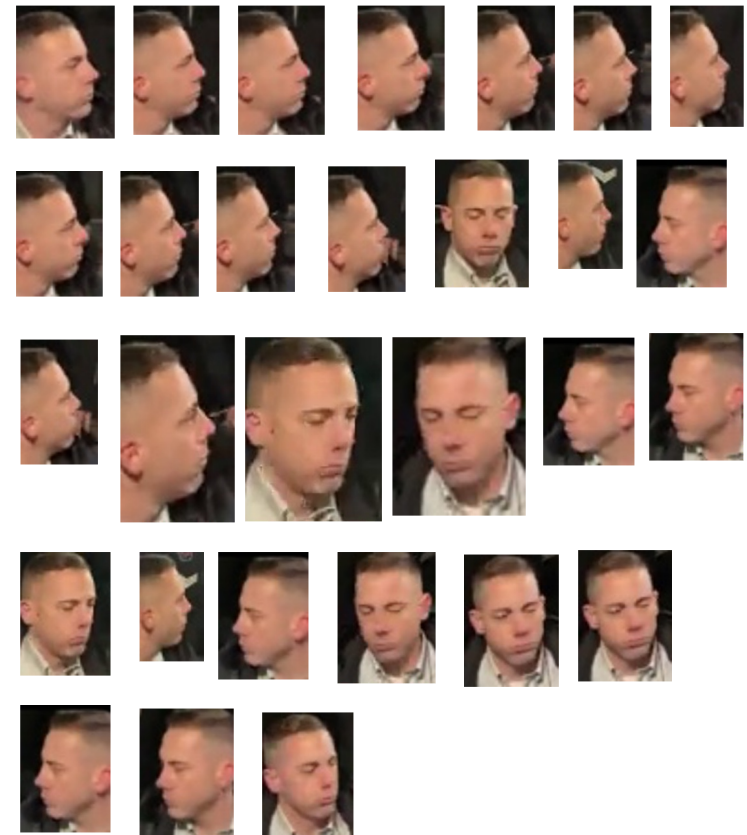
- Group visually very similar frames and process the least blurry from each group
- Fragment video to shots and subshots
- **Detect key elements (faces, text)**
- Group similar key elements by tracking the motion of the detections' bounding box
- Estimate the blurriness of each key element in the group and select the sharpest item
- Employ super-resolution techniques to enhance the selected item



AI-powered keyframe selection and enhancement Workflow



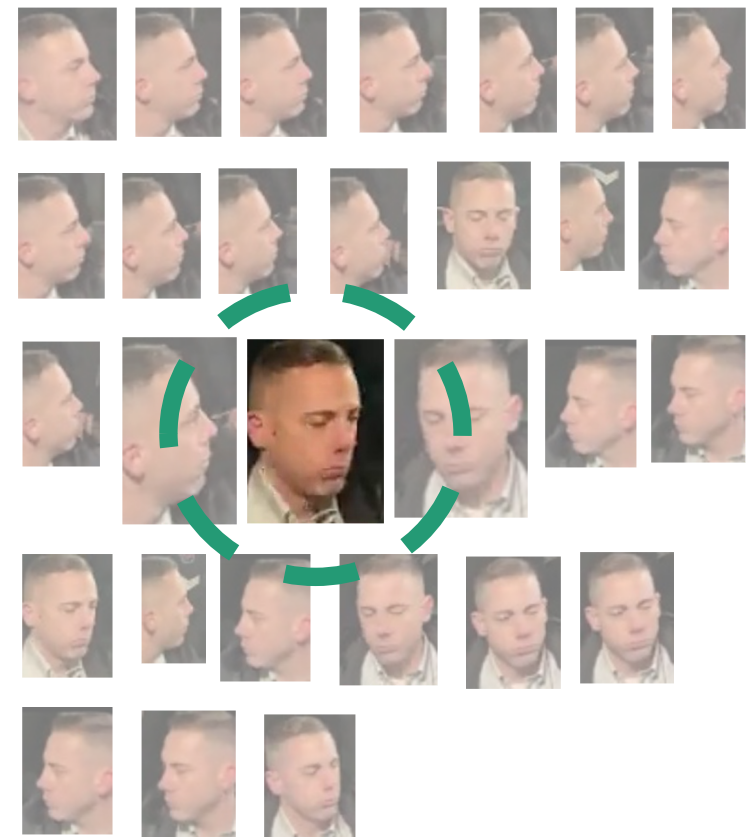
- Group visually very similar frames and process the least blurry from each group
- Fragment video to shots and subshots
- Detect key elements (**faces**, text)
- Group similar **face** elements by tracking the motion of the detections' bounding box
- Estimate the blurriness of each **face** element in the group and select the sharpest item
- Employ super-resolution techniques to enhance the selected item



AI-powered keyframe selection and enhancement Workflow



- Group visually very similar frames and process the least blurry from each group
- Fragment video to shots and subshots
- Detect key elements (**faces**, text)
- Group similar **face** elements by tracking the motion of the detections' bounding box
- Estimate the blurriness of each **face** element in the group and select the sharpest item
- Employ super-resolution techniques to enhance the selected item



AI-powered keyframe selection and enhancement Workflow



- Group visually very similar frames and process the least blurry from each group
- Fragment video to shots and subshots
- Detect key elements (**faces**, text)
- Group similar **face** elements by tracking the motion of the detections' bounding box
- Estimate the blurriness of each **face** element in the group and select the sharpest item
- Employ super-resolution techniques to enhance the selected item



AI-powered keyframe selection and enhancement Workflow



- Group visually very similar frames and process the least blurry from each group
- Fragment video to shots and subshots
- Detect key elements (faces, **text**)
- Group similar **text** elements by tracking the motion of the detections' bounding box
- Estimate the blurriness of each **text** element in the group and select the sharpest item
- Employ super-resolution techniques to enhance the selected item



AI-powered keyframe selection and enhancement Workflow



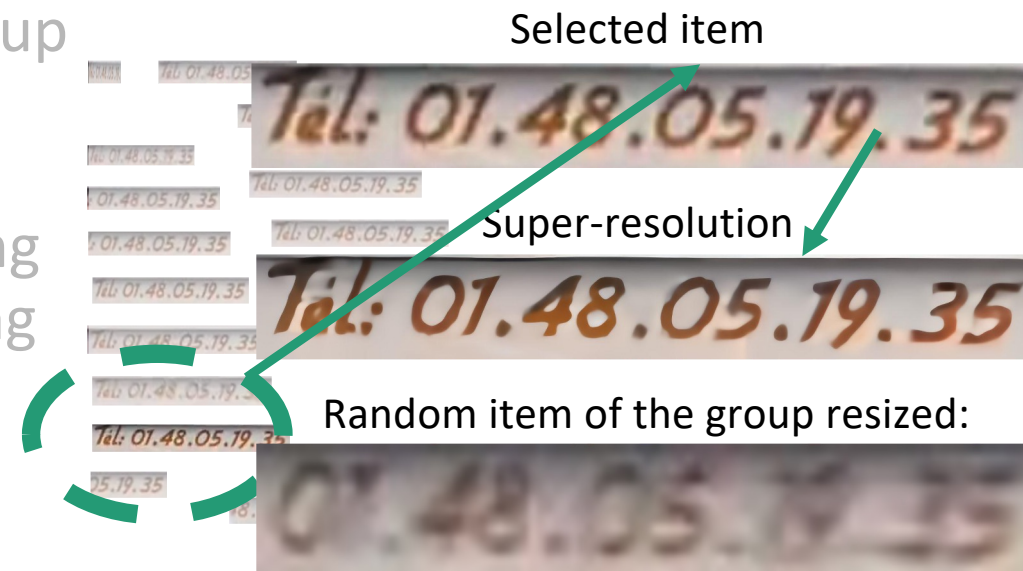
- Group visually very similar frames and process the least blurry from each group
- Fragment video to shots and subshots
- Detect key elements (faces, **text**)
- Group similar **text** elements by tracking the motion of the detections' bounding box
- Estimate the blurriness of each **text** element in the group and select the sharpest item
- Employ super-resolution techniques to enhance the selected item



AI-powered keyframe selection and enhancement Workflow



- Group visually very similar frames and process the least blurry from each group
- Fragment video to shots and subshots
- Detect key elements (faces, **text**)
- Group similar **text** elements by tracking the motion of the detections' bounding box
- Estimate the blurriness of each **text** element in the group and select the sharpest item
- Employ super-resolution techniques to enhance the selected item





The KSE service

AI-powered keyframe selection and enhancement

The KSE service



- Encapsulated all discussed technologies in the Keyframe Selection and Enhancement (KSE) service
- This is a web service that automates keyframe selection and enhancement to help users quickly verify video authenticity
- Simplicity: Web-Based - no installation; integrates with workflows

AI-powered keyframe selection and enhancement

Methods employed



- Instead of relying solely on benchmark results from papers, we tested several recent state-of-the-art methods on in-house data and measured their performance on the same hardware.

Label	Preview	Visual inspection score ↑	No of params	Inference time	LPIPS score ↓	SSIM score ↑	PSNR score ↑	FID (Inception) score ↓	FID (ViT) score ↓	FID (MobileNet) score ↓	KD applica
GT		1.00			0.000	1.000	999999999	0.000	0.289		0
Baseline (resize)		0.40			0.272	0.815	28.032	281.001	181.693	1158.028	
06 VQFR		0.65	83,486,539	0.405	0.305	0.690	23.317	317.040	318.189	1294.306	
07 RestoreFormer++		0.80	79,308,423	0.281	0.297	0.738	25.696	347.200	336.081	1234.799	
09 DiffFace		0.90	159,788,454	16.807	0.246	0.778	27.167	163.450	218.194	363.885	

AI-powered keyframe selection and enhancement

Methods employed



Task	Method	Score
Video shot fragm.	TransNet v2	77.9% F1 score on ClipShots (2nd place)
Video subshot fragm.	MDAFGTS	2nd in Annotated-dataset-sub-shot-segmentation
Face detection		96.175 mAP on WiderFace (2nd place)
	RetinaFace	93.12 mAP on WiderFace (much faster than Retina Face)
	Yolo V7 FD	
Text detection	MixNet	90.5% F-score on TotalTest (best score)
	FAST	86.4% F-score on TotalTest (8th place)
Multi-object tracking	Deep OC-SORT	64.9% HOTA on MOT17 (SoA 66.6%)
	OC-SORT	63.2% HOTA on MOT17 (100x faster than SoA)
Face super-resolution	CodeFormer	26.49 PSNR on CelebA-HQ (2nd place)
	GCFSR	22.18 PSNR on CelebA-HQ (more realistic results)
Text super-resolution	Telescope	46.34% average accuracy on TextZoom
General super-resolution	SwinSR V2	24.48 PSNR on Set5A (best performing)

AI-powered keyframe selection and enhancement

Methods employed



Task	Method	Score
Video shot fragm.	TransNet v2	77.9% F1 score on ClipShots (2nd place)
Video subshot fragm.	MDAFGTS	2nd in Annotated-dataset-sub-shot-segmentation
Face detection	RetinaFace	96.175 mAP on WiderFace (2nd place)
	Yolo V7 FD	93.12 mAP on WiderFace (much faster than Retina Face)
Text detection	MixNet	90.5% F-score on TotalTest (best score)
	FAST	86.4% F-score on TotalTest (8th place)
Multi-object tracking	Deep OC-SORT	64.9% HOTA on MOT17 (SoA 66.6%)
	OC-SORT	63.2% HOTA on MOT17 (100x faster than SoA)
Face super-resolution	CodeFormer	26.49 PSNR on CelebA-HQ (2nd place)
	GCFSR	22.18 PSNR on CelebA-HQ (more realistic results)
Text super-resolution	Telescope	46.34% average accuracy on TextZoom
General super-resolution	SwinSR V2	24.48 PSNR on Set5A (best performing)

AI-powered keyframe selection and enhancement

The KSE service



Older service of the plugin:

- Video fragmentation to subshots
- Selects a pre-specified number of (default is 3) keyframes from each subshot

Updates within vera.ai:

- Video fragmentation to shots, and subshots
- Automatically calculates the number of keyframes required to adequately describe the content of each subshot
- Detection of key elements (i.e. faces and text regions, for now)
- Grouping of similar key elements in the same shot
- Enhancement of key elements

The KSE service in the verification plugin

The KSE service replaced the existing video fragmentation service in the verification plugin

- The video fragmentation service was very fast!
- The video fragmentation has a user-base
- KSE must maintain a similar execution speed despite its added complexity

AI-powered keyframe selection and enhancement

Performance Optimization



- Group visually very similar frames and process the least blurry from each group
- Fragment video to shots and subshots
- Detect key elements (faces, text)
- Group similar key elements by tracking the motion of the detections' bounding box
- Estimate the blurriness of each key element in the group and select the sharpest item
- Employ super-resolution techniques to enhance the selected item

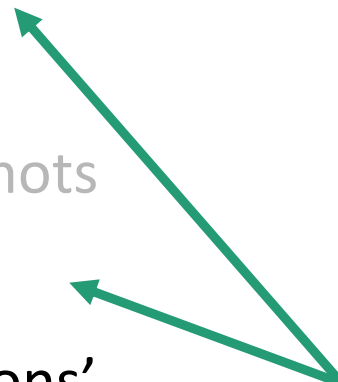
AI-powered keyframe selection and enhancement

Performance Optimization



- Group visually very similar frames and process the least blurry from each group
- Fragment video to shots and subshots
- Detect key elements (faces, text)
- Group similar key elements by tracking the motion of the detections' bounding box
- Estimate the blurriness of each key element in the group and select the sharpest item
- Employ super-resolution techniques to enhance the selected item

Stages for speeding up the analysis process



AI-powered keyframe selection and enhancement

Performance Optimization Results



Video duration (seconds)	991	238	578	1193	60
Video fragmentation					
(process time seconds)	48	21	46	77	5
(process time % to video duration)	4.84%	8.82%	7.96%	6.45%	8.33%
KSE					
(process time seconds)	148	165	255	857	47
(process time % to video duration)	14.93%	69.33%	44.12%	71.84%	78.33%

AI-powered keyframe selection and enhancement

Performance Optimization Results



Video duration (seconds)	991	238	578	1193	60
Video fragmentation					
(process time seconds)				77	5
(process time % to video duration)					%
KSE					
(process time seconds)	148	165	255	857	47
(process time % to video duration)	14.93%	69.33%	44.12%	71.84%	78.33%

KSE offering a much wider spectrum of functionalities while still remaining real-time

The KSE service in the verification plugin



The KSE service replaced the existing video fragmentation service in the verification plugin

The KSE service in the verification plugin



The KSE service replaced the existing video fragmentation service in the verification plugin



<https://www.invid-project.eu/tools-and-services/invid-verification-plugin/>



The standalone KSE service

The KSE service has also a standalone UI which you can freely use!

Some limitations:

- Video duration must be under 10 minutes
- If local video file is submitted, the filesize must be under 200MB

The standalone KSE service

The KSE service has also a standalone UI which you can freely use!

Some limitations:

- Video duration must be under 10 minutes
- If local video file is submitted, the filesize must be under 200MB



<https://kse.idt.itl.gr>



Thank you!

This work was supported by the EU Horizon Europe programme under grant agreement 101070093 vera.ai

Contact points
kapost@iti.gr
bmezaris@iti.gr

Kostas Apostolidis / Research Assistant at CERTH

Contact: bmezaris@iti.gr, kapost@iti.gr



CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS



ATHENS TECHNOLOGY CENTER



**Deutsche
Welle**



UNIVERSITÀ DEGLI STUDI
DI NAPOLI FEDERICO II



1506
UNIVERSITÀ
DEGLI STUDI
DI URBINO
CARLO BO



UNIVERSITEIT
VAN AMSTERDAM



université
PARIS-SACLAY



OPERATING EUROVISION AND EURORADIO



Follow us on Twitter: [@veraai_eu](https://twitter.com/veraai_eu)

Website: <https://www.veraai.eu/>

Co-financed by the European Union, Horizon Europe programme, Grant Agreement No 101070093.

Additional funding from Innovate UK grant No 10039055 and the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract No 22.00245

