A large-Scale TV Dataset for partial video copy detection

PhD student
Van-Hao Le[1]

Supervisors
Donatello Conte[1]
Mathieu Delalandre[1]

[1]LIFAT Laboratory, Tours city, France

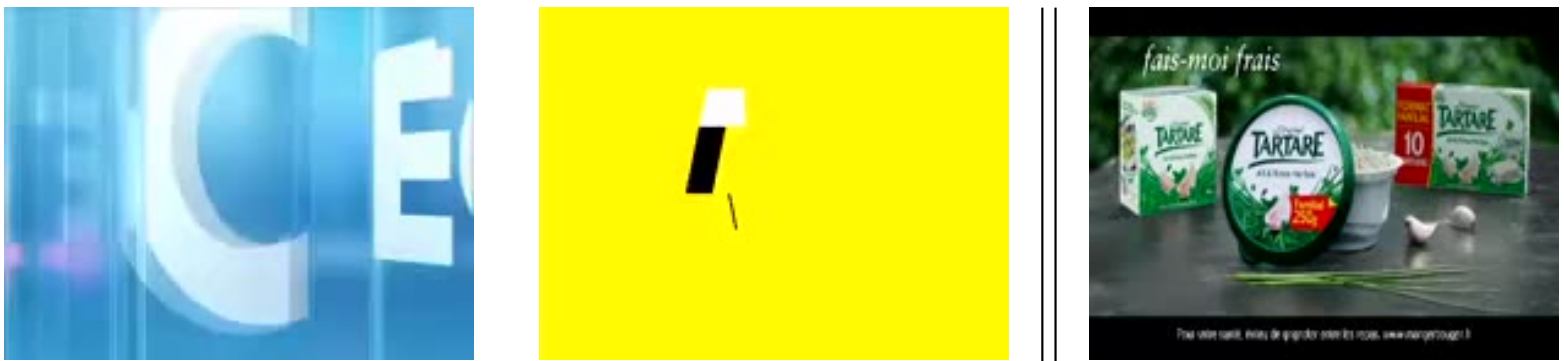
July 12th, 2022

Outline

1. Introduction
2. Our approach
   - The TV protocols
   - Our protocol
   - A large-Scale TV Dataset (STVD)
3. Performance Evaluation
4. Conclusions
5. Perspectives

Introduction (1/3)

- Partial video copy detection (PVCD) aims at finding short segment(s) which have transformed in long video(s).

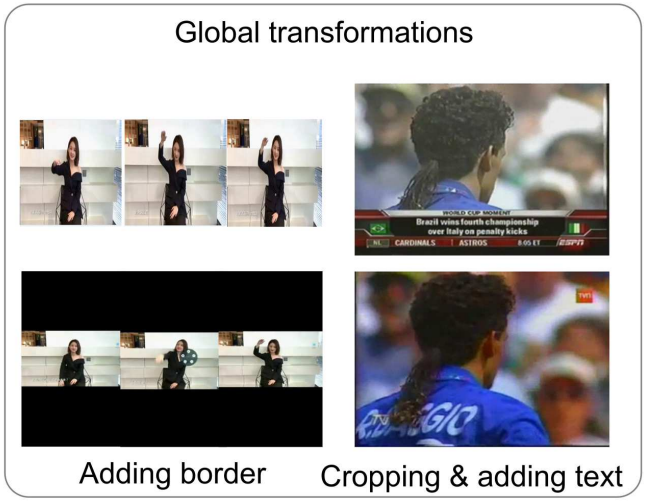


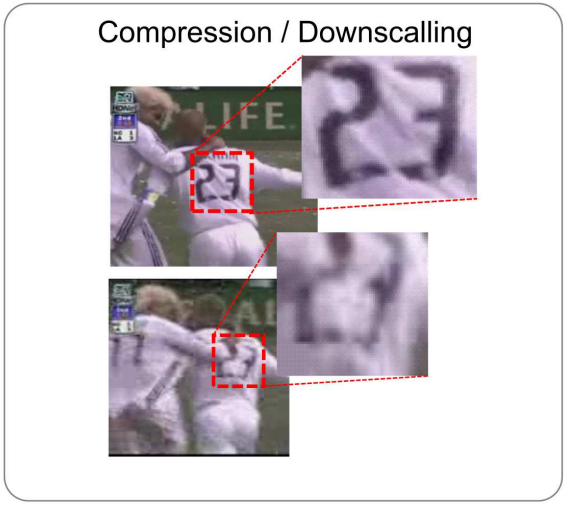
short segments | a long video

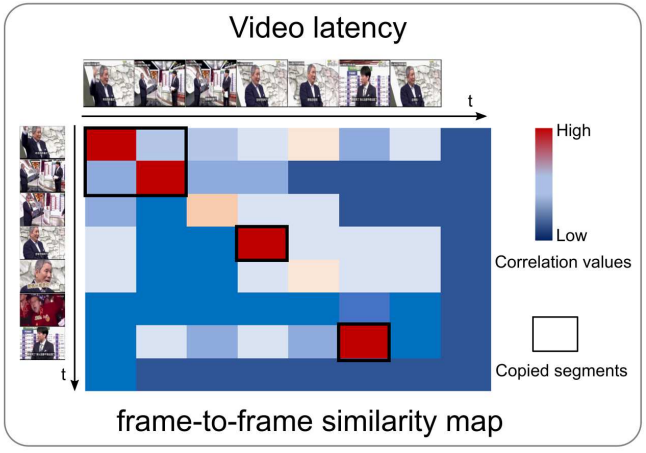
- PVCD includes several application domains (copyright protection, video retrieval, etc.).
- It is a key topic in the Computer Vision field [Guzman, 2019; Liu, 2021; Tan, 2022].

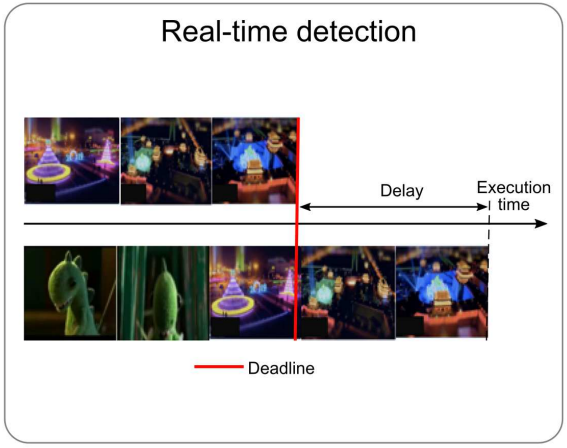
Introduction (2/3)

- PVCD addresses different detection problems [Kordopatis, 2017; Liu, 2021; Tan, 2022].









Introduction (3/3)

- Needs of public datasets for performance evaluation
  - 4 datasets have been published with protocols.
  - Protocols process mainly from Web videos with real degradations and manual annotation.

Datasets	CC_WEB [Wu, 2007]	SVD [Jiang, 2019]	VCDB [Jiang, 2014]	VCSL [He, 2022]
No longer in use[1]	✓			
Small size	✓	✓		
Unbalanced data			✓	
Costly annotation[2]		✓	✓	✓
None / Little control of degradations	✓	✓	✓	✓
None frame-level annotation	✓	✓	✓	✓

[1] The mAP score of CC_WEB reached at $\simeq$ 0.976 [Kordopatis, 2017]

[2] About $\simeq$ 700 man-hours for labeling 528 positive & 100,000 negative videos in VCDB

The TV protocols (1/2)

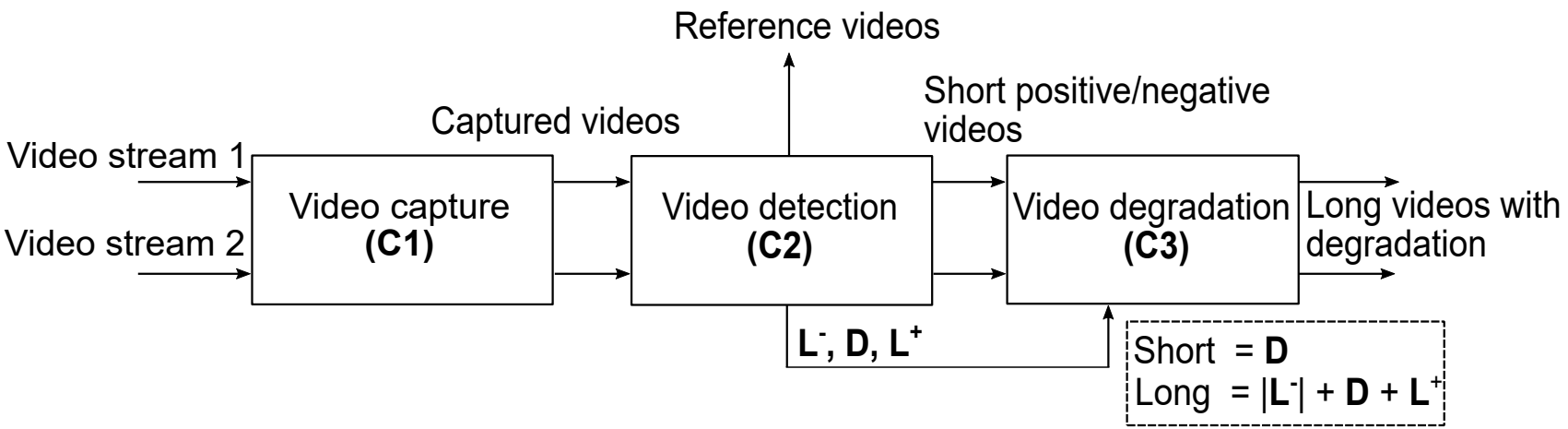
- New datasets must be proposed for performance characterization:
  - scalable with balanced data,
  - applicable to several performance characterization tasks,
  - with a frame-level annotation for timestamping.
- An alternative is to process with TV-based protocols.

The TV protocols (2/2)

- 3 proposed protocols with datasets [Joly, 2007; Law-To, 2007; Chenot, 2014], not public available. (see Appx.)
- The protocols used to design the datasets:
  - separate positive / negative captures,
  - apply synthetic generation of partial copies,
  - use a full-search strategy to detect real copies.
- We propose a new protocol able to:
  - separate captures similar to [Joly, 2007; Law-To, 2007],
  - avoid a full-search strategy with a metadata support,
  - extract real / true-life partial copies,
  - have a fine control of degradations.

Our protocol (1/5)

- The system architecture includes 3 main components:
  - **(C1)** captures videos with a TV workstation,
  - **(C2)** annotates reference, positive and negative videos,
  - **(C3)** generates test sets with synthetic degradations.



(see Appx.)

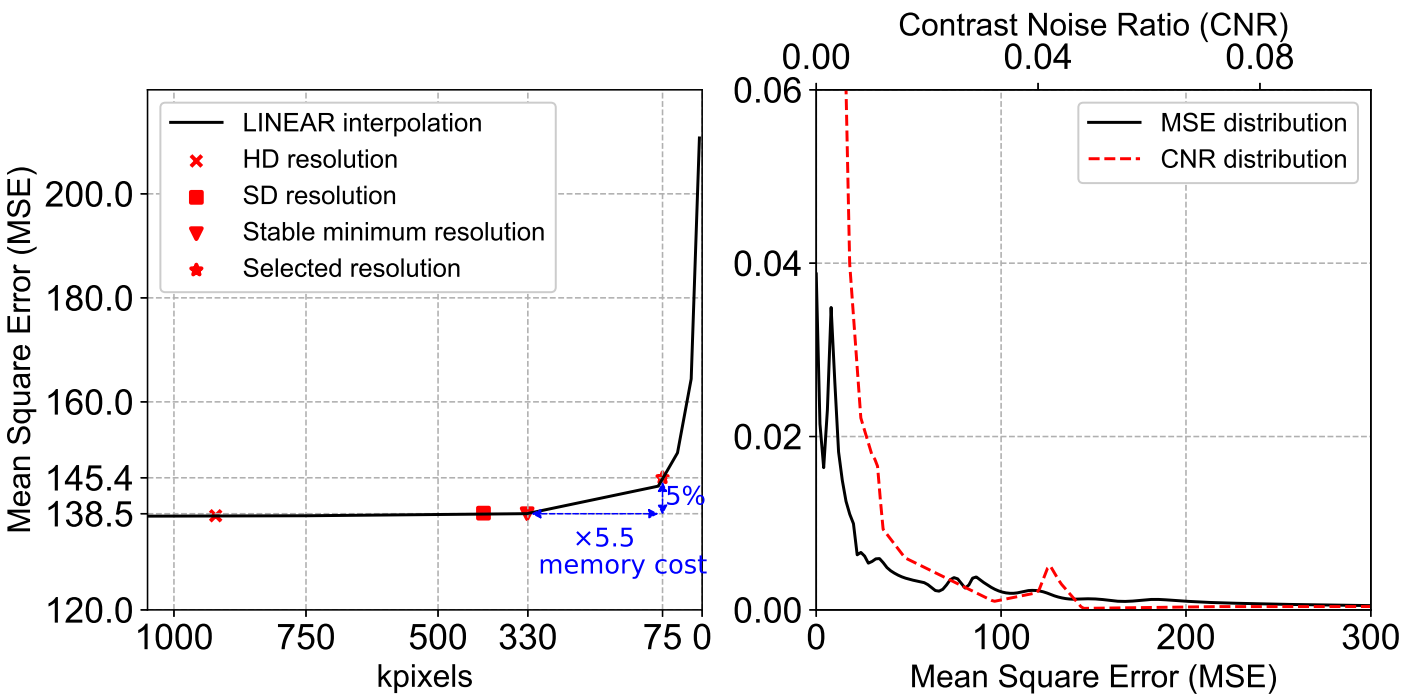
Our protocol (2/5)

Video capture **(C1)**

- Capture French DTT using a TV workstation setup:

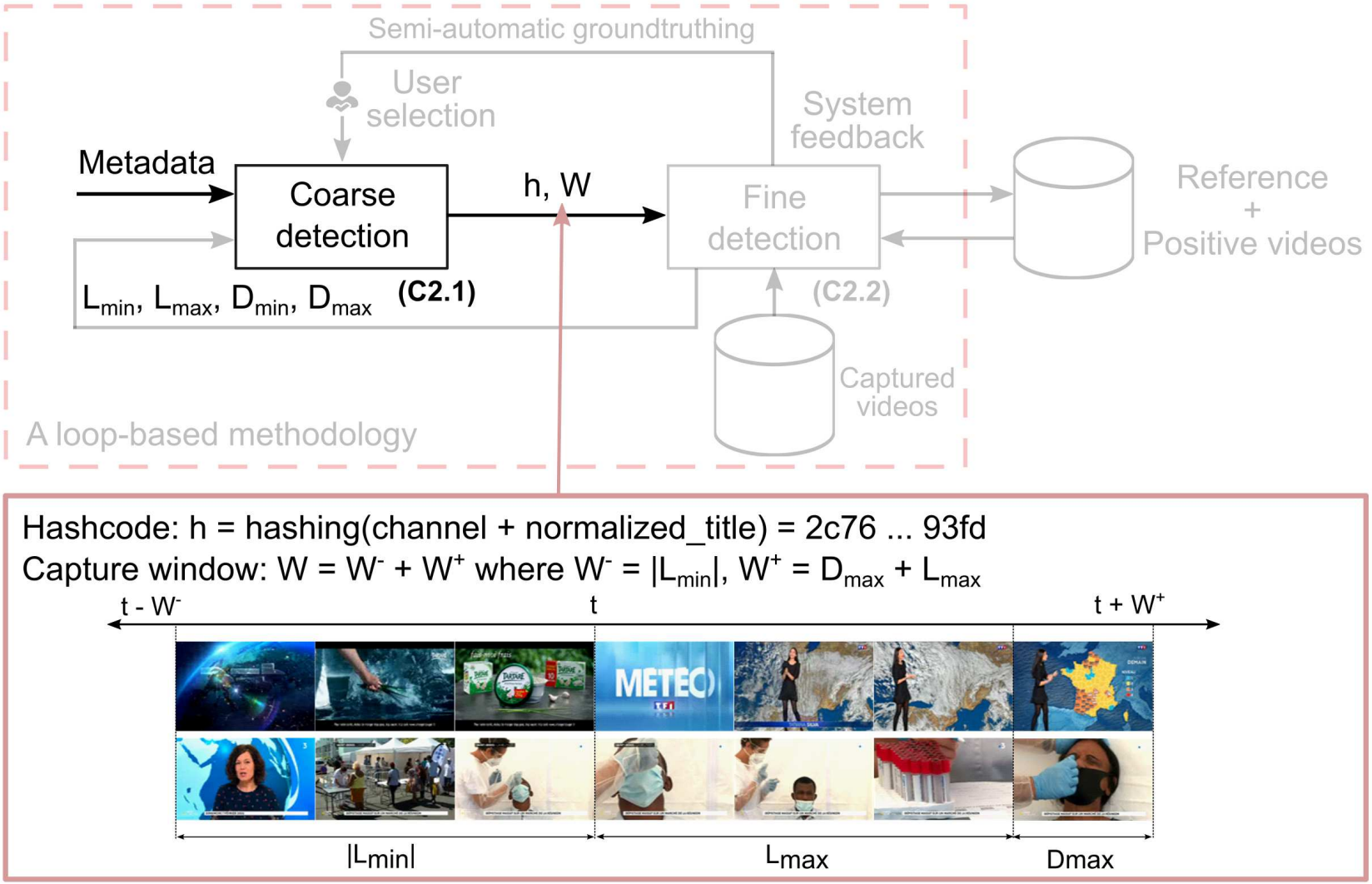
Channels per month	Daily file					Total size (TB)	
	Resolution	kbps	Aspect ratio	Length	FPS	Files	Size
8	320 × 240	560	4 : 3	20 h	30	720	3.46

- An optimized strategy of memory cost and a noise level



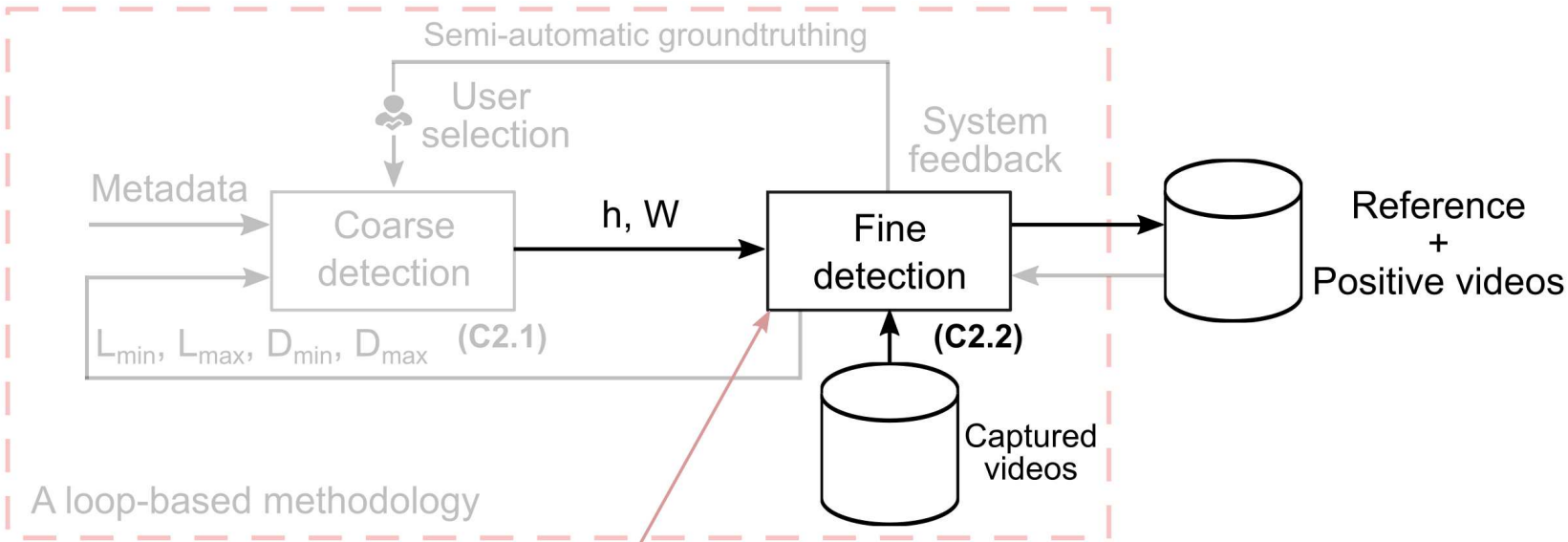
Our protocol (3/5)

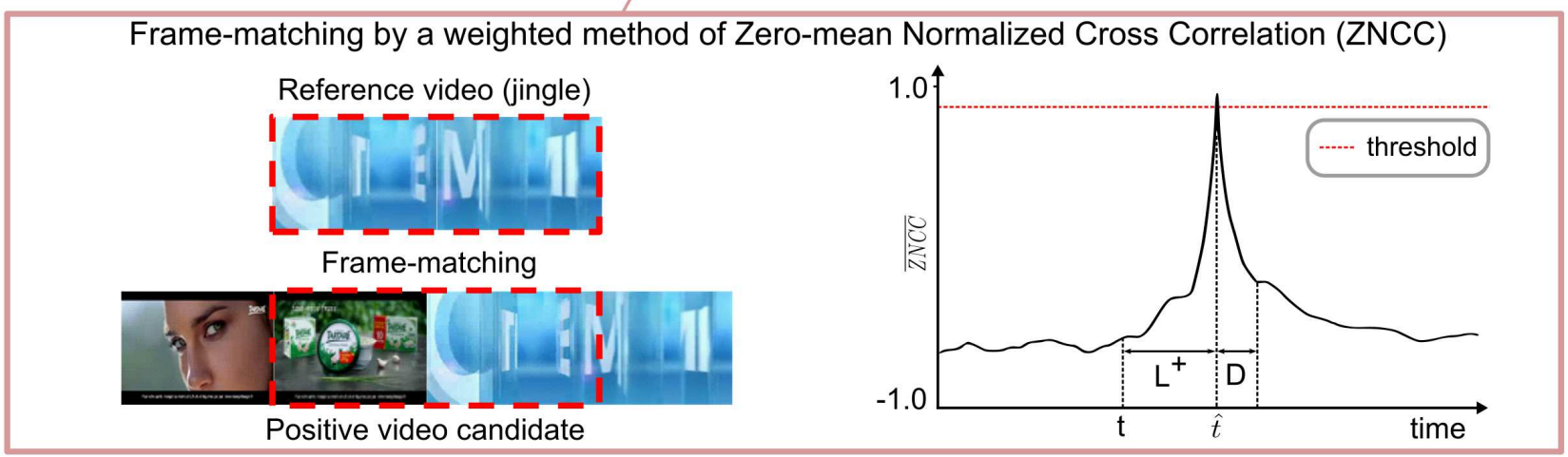
Video detection (C2)



Our protocol (3/5)

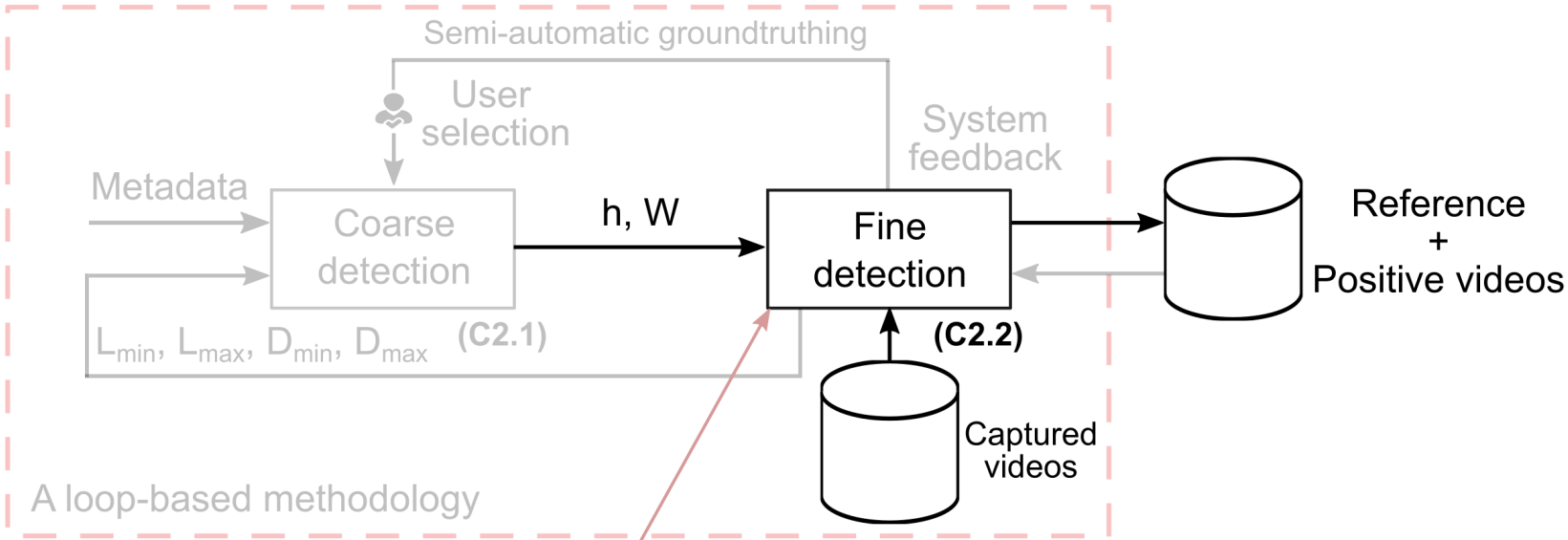
Video detection (C2)



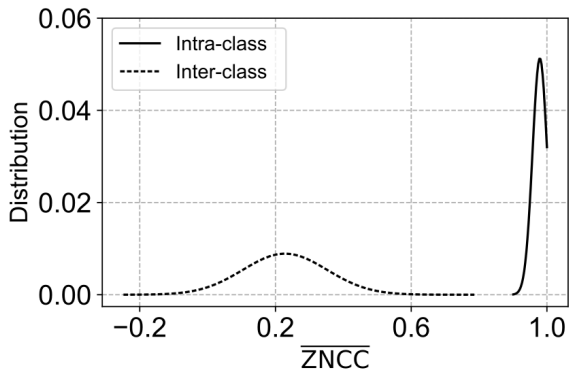


Our protocol (3/5)

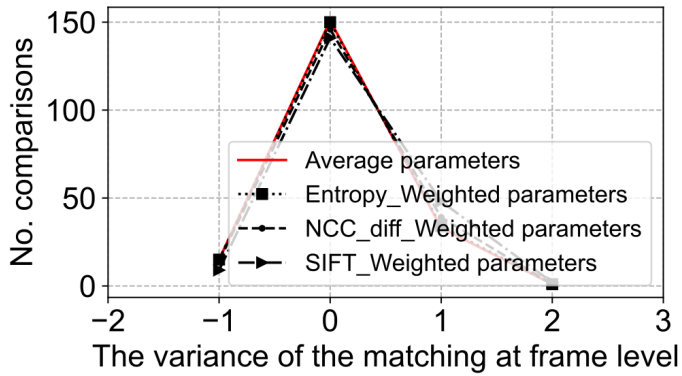
Video detection (C2)



The separation between the inter-class and intra-class distribution

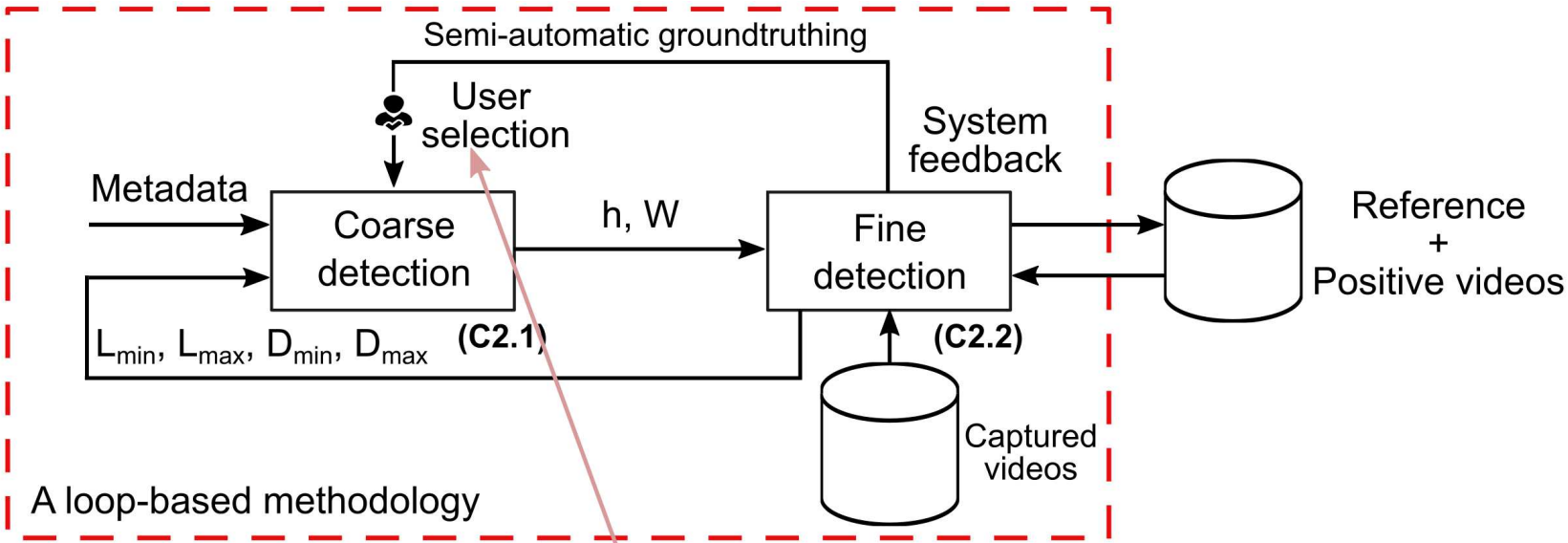


The accuracy of weighting methods for timestamping



Our protocol (3/5)

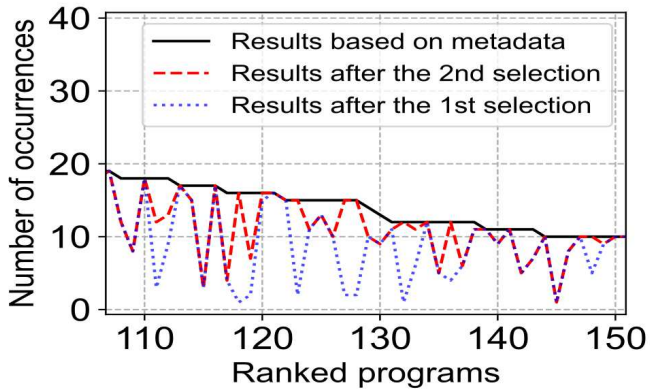
Video detection (C2)



Some error-prone cases

Different content Near-duplicate

Cost minimization for user interaction

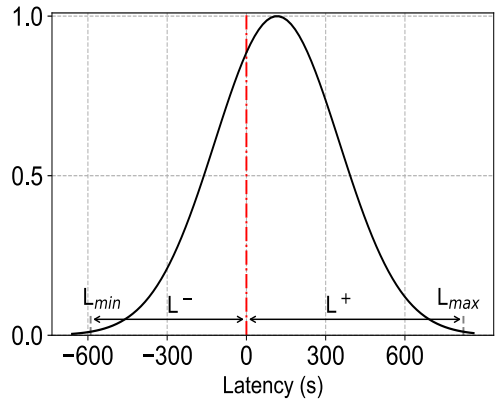


Our protocol (4/5)

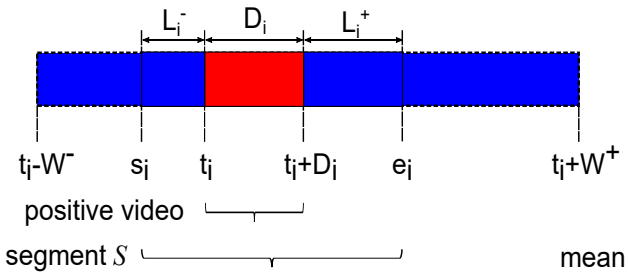
Video degradation **(C3)**

- **(C2)** extracts real-life partial video copies:
  - applying the latency model detected with **C2**,
  - detecting the overlapping cases for merging.

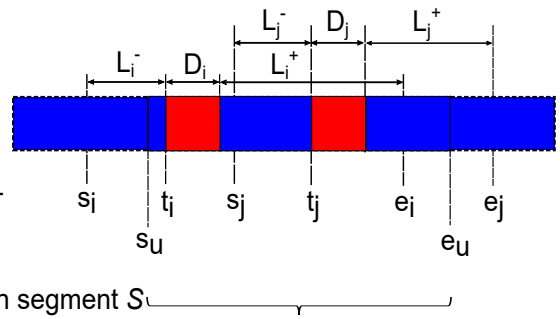
An approximately Gaussian distribution



A normal case of a partial copy



An overlapping case



(see Appx.)

Our protocol (5/5)

Video degradation **(C3)**

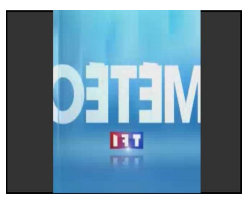
- The final dataset with degradation methods: 6 sets (A-F)

	Test set	Video cut	Downscaling	Compression	Flipping	Rotating	Black-border	Video speeding
		T_0	T_1	T_2	T_3	T_4	T_5	T_6
A	Root capture	✓						
B	'Hello World'	✓	✓	✓				
C	Pixel attack	✓	✓	✓				
D	Global transformations	✓	✓	✓	✓	✓	✓	
E	Video speeding	✓	✓	✓				✓
F	Combination	✓	✓	✓	✓	✓	✓	✓

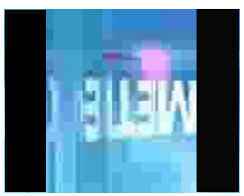
Set A

Set B

Set C


Set D


Set E


Set F

A large-Scale TV Dataset (STVD)

- STVD statistics: included 6 test sets in **C3**

	(C1)		(C2)		(C3)	
	Channels	Duration	Videos	Duration	Videos	Duration
Positive videos	8	4,800 h	3,780	6 h	19,280	2,515 h
Negative videos	16	9,600 h	12,165	21 h	64,040	8,145 h

- The comparisons between STVD versus VCDB, VCSL

Datasets	**VCDB** [Jiang, 2014]	**VCSL** [He, 2022]	**STVD** Ours
Reference videos	28	122	243
Positive videos	528	9,207	19,280
Negative videos	100,000	N/A	64,040
Duration (h)	2,000	N/A	10,660
Pairs of partial copies	9,073	276,303	418,782
Noise characterization	real noise	real noise	noise-free
Annotation cost (m-h)	700	20,000	105
Frame-level annotation	✗	✗	✓

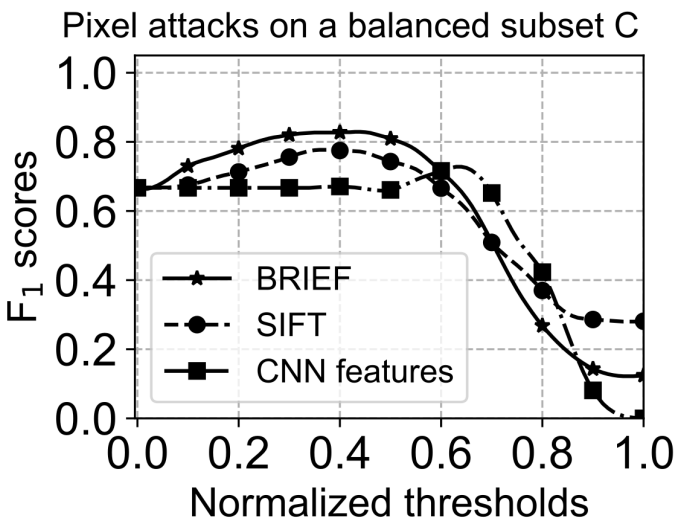
(h): hours, (m-h): man-hours and N/A: not available

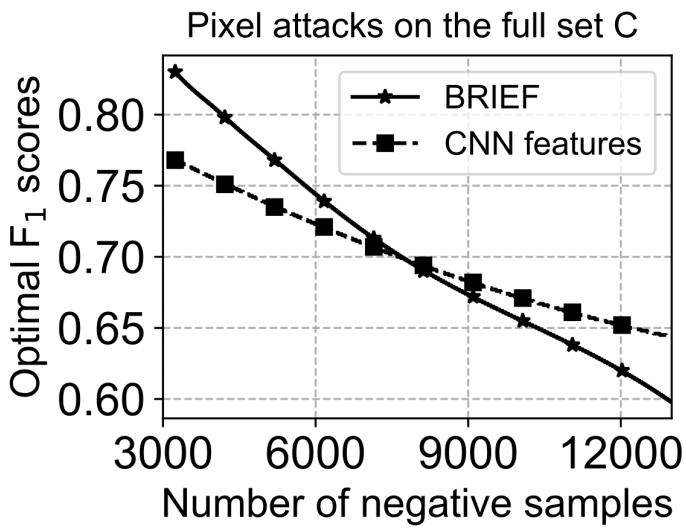
Performance Evaluation

- Baseline methods for performance evaluation

Method	Key-frame extraction	Frame matching
[Zhu, 2016]	TC-SIFT features	K-NN searching with LSH
[Zhang, 2016]	FPS sampling & CGM	Matching with BRIEF features
[Zhang, 2020]	Frame clustering	Matching with CNN features

- Performance characterization for detection
  - Metrics: Precision, Recall, F-measure,
  - Test set: B for a 'Hello world' ability with $F_1 \simeq 0.98$,
  - Test set: C for pixel attacks.



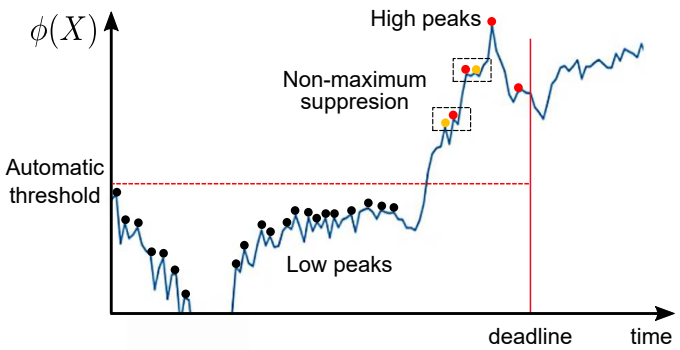


Conclusions

- We proposed a protocol to design a dataset for PVCD:
  - ensuring the dataset scalability with balanced data,
  - offering a fine control of degradation,
  - able to annotate at a frame-level for timestamping.
- We published a large-scale dataset for PVCD:
  - $\simeq$420K pairs, $\simeq$83K videos, 10K hours,
  - with a baseline comparison showing room for improvement,
  - The dataset is public available at https://dataset-stvd.univ-tours.fr/(PVCD)
- Our protocol can be adapted to other problems:
  - computer vision / fact checking (see Appx.),
  - operational research / parallel machine scheduling.

Perspectives (1/2)

- Real-time PVCD processes with a deadline (e.g., 1-3s).
  - There are few contributions in the literature [Guzman, 2016; Tan, 2022].
- We have designed a new approach processing:
  - with a workstation supporting a real-time video decoding and ZNCC, (see Appx.),
  - with a key-frame selection using a goodness criterion.



Reference	Method	Δ (ms)	F_1 (%)	#KF
[Özbulak, 2016]	SURF		92.71	1356
[Zhang, 2016]	BRIEF		92.82	1353
[Guzman, 2016]	NCC difference	1000	93.84	1350
[Hou, 2018]	Max entropy		97.71	1353
[Zhang, 2020]	CNN features		98.25	1351
Ours	Real-time NCC	1000	99.11	1351

- Our results outperformed on a subset SVD (22 h).
  - Scalability must be investigated (STVD is $\times 80$ bigger),
  - For hard/severe degradations, real-time deep learning will matter [Daghaghi, 2021; Liang, 2021].

Perspectives (2/2)

- Performance evaluation on PVCD:
  - first experiments show room for improvement,
  - STVD must be opened to the research community (direct partnership[3], research mailing lists[4], contests[5]),
  - new protocols/metrics should be investigated [He, 2022].
- Trends for robust PVCD is to investigate deep learning:
  - embedding the spatial-temporal information within the model [Hu, 2019; Han, 2021; Liu, 2021],
  - mixing deep learning with temporal networks [Tan, 2022],
  - STVD could help to design new strategies / models with a fine control of degradation and scalability for training.

[3] List of works on PVCD: [Guzman, 2019; Hu, 2019; Han, 2021; Liu, 2021; Tan, 2022;]

[4] tout-isis@lists.gdr-isis.fr , cvml@lists.auth.gr , connectionists@mailman.srv.cs.cmu.edu

[5] ICIAP competition 2021, ECCV competition 2022, CVPR Challenge 2022

List of publications

1. F. Rayar, M. Delalandre and **V.H. Le**. *A large-scale TV video and metadata database for French political content analysis and fact-checking*. Conference on Content-Based Multimedia Indexing (CBMI), 2022[6].
2. **V.H. Le**, M. Delalandre and D. Conte. *A large-Scale TV Dataset for partial video copy detection*. International Conference on Image Analysis and Processing (ICIAP), Lecture Notes in Computer Science (LNCS), vol 13233, pp. 388-399, 2022.
3. **V.H. Le**, M. Delalandre and D. Conte. *Une large base de données pour la détection de segments de vidéos TV*. Journées Francophones des Jeunes Chercheurs en Vision par Ordinateur (ORASIS), 2021.
4. **V.H. Le**, M. Delalandre and D. Conte. *Real-time detection of partial video copy on TV workstation*. (CBMI), pp. 1-4, 2021.

[6] The paper was accepted for presentation in September, 2022.

Thank you for your attention !

Appendix - PVCD datasets (1/2)

- The state-of-the-art datasets

Datasets	Year	Reference videos	Positive videos	Negative videos	Duration (h)	Pairs of partial copies	Public available
TV_2007	2007	100	500	N/A	60,000	N/A	No
BBC_2007	2007	9	9	N/A	3,1	N/A	No
CC_WEB	2007	24	3,481	9,309	537	N/A	Yes
TRECVID	2010	1,608	134	7,866	200	N/A	No
TV_2014	2014	N/A	20,000,000	N/A	380,000	N/A	No
VCDB	2014	28	528	100,000	2,000	9,073	Yes
SVD	2019	1,026	10,211	26,927	2,705	N/A	Yes
VCSL	2022	122	9,207	N/A	N/A	276,303	Yes
STVD	2021	243	19,280	64,040	10,660	418,782	Yes

(h): hours and N/A: not available

Appendix - PVCD datasets (2/2)

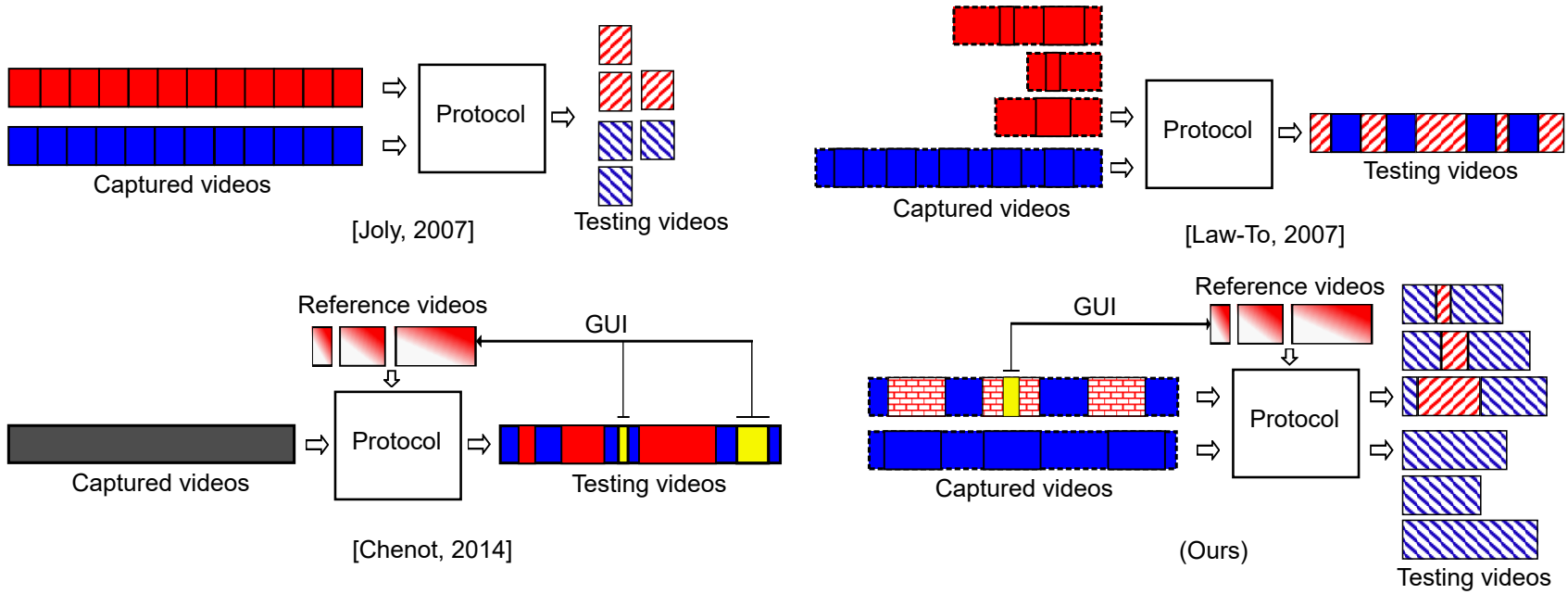
- The state-of-the-art datasets

Datasets	Source of capture	Annotation cost(m-h)	Degradation methods	Groundtruth level	Top F1 score
TV_2007	TV	N/A	synthetic	video	0.840
BBC_2007	TV	N/A	synthetic	video	0.860
CC_WEB	Web	N/A	real	video	0.980
TRECVID	Web	N/A	synthetic	segment	0.795
TV_2014	TV	N/A	synthetic	video	N/A
VCDB	Web	700	real	segment	0.876
SVD	Web	800	synthetic	video	0.780
VCSL	Web	20,000	real	segment	0.874
STVD	TV	105	synthetic	frame	N/A

(m-h): man-hours and N/A: not available

Appendix - the TV protocols (1/2)

- The protocols proposed in the literature.

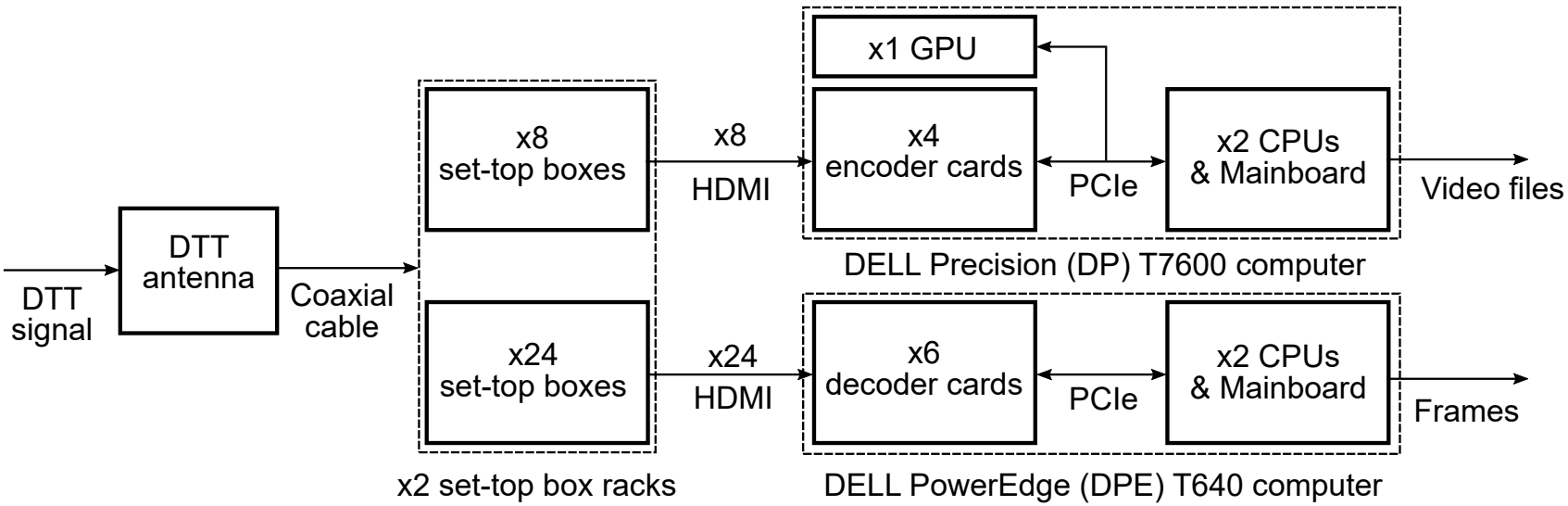


Appendix - the TV protocols (2/2)

- The process of the negative videos extraction:
  1. separates the captures,
  2. extracts timestamp t for TV programs,
  3. makes idle/not use of surrounding t with the window,
  4. splits into successive intervals for not idle sequences.

Appendix - the TV workstation

- The architecture of the TV workstation



- Main hardware specification information

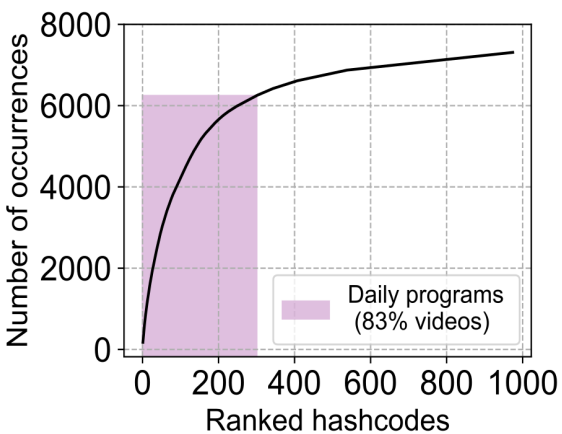
Name	Model	Qty	Specification
Computer DP	DELL Precision T7600	1	×2 Intel Xeon(R) CPU E5-2620, 64 GB RAM
Computer DPE	DELL PowerEdge T640	1	×2 Intel Xeon Gold 5218R CPU 2.10, 32 GB RAM
Encoding card	Avermedia CL332-HN	4	2-Channel Full HD HW H.264
Decoding card	Avermedia CE314-HN	6	4-Channel Full HD
Set-top box	Astrell DVB-T 011128	32	Adapter DVB-T HDMI
GPU	GeForce RTX2070 super	1	9.06 TFLOPS SP

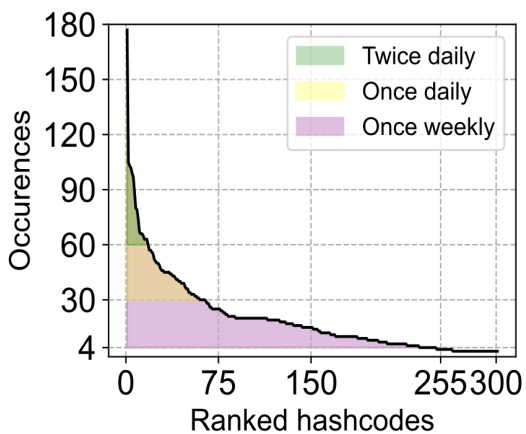
Appendix - the (C2) component

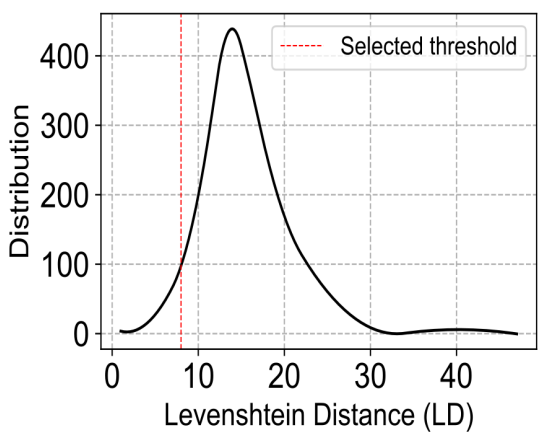
- The user optimization at the meta-data level

Channel	Programs were removed			Programs were kept			LD
	Normalized title	Hashcode	Videos	Normalized title	Hashcode	Videos	
M6	le1245	e38 . . . 302	5	le1945	f2b . . . f58	19	1
TF1	swat	020 . . . 2f5	9	swat2017	648 . . . a64	15	4

- The metadata processing with the NLP method







Appendix - test sets detail

- The degradation methods used to generate the STVD.

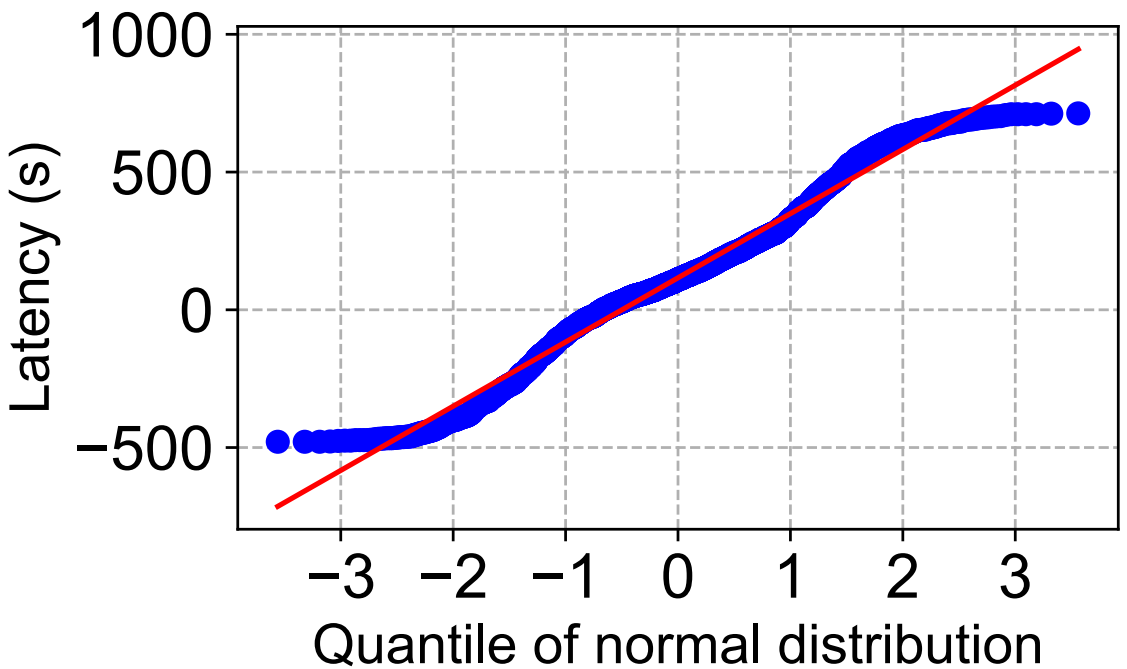
Label	Method	Parameters
T_0	video cut	introduced latency $\lvert \mathbf{L}^- \rvert, \mathbf{L}^+$
T_1	down-scaling	$\alpha \in [0.1, 0.9]$
T_2	compression	video bitrate $\frac{1}{\beta}$ with $\beta \in [1, 80]$ kbps
T_3	flipping	applies randomly (yes/no)
T_4	rotating	rotation $\in \{0, \frac{\Pi}{2}, \Pi, \frac{3}{2}\Pi\}$
T_5	black border & stretching	aspect ratio $\frac{w}{h} \in \{0.46, 1.78, 2.17, etc.\}$
T_6	video speeding	FPS $\in [15, 25]$

- 6 test sets are created considering different aspects.

Test set	T_0	T_{1-2}	$\alpha \in$	$\beta \in$	T_{3-5}	T_6	Description
Set A	$\surd$						Root capture for tuning
Set B	$\surd$	$\surd$	$[0.25, 0.9[$	$[1, 40[$			"Hello world" test set
Set C	$\surd$	$\surd$	$[0.1, 0.25]$	$[40, 80]$			Pixel attack with scalability
Set D	$\surd$	$\surd$	0.6	20	$\surd$		Global transformations with scalability
Set E	$\surd$	$\surd$	0.6	20		$\surd$	Video speeding with scalability
Set F	$\surd$	$\surd$	$[0.1, 0.25]$	$[40, 80]$	$\surd$	$\surd$	Combination of sets C, D and E

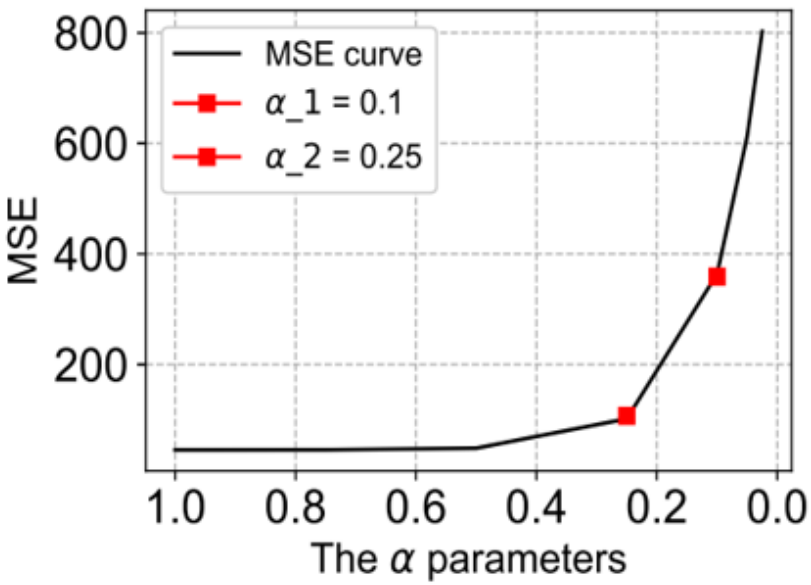
Appendix - the Gaussian model

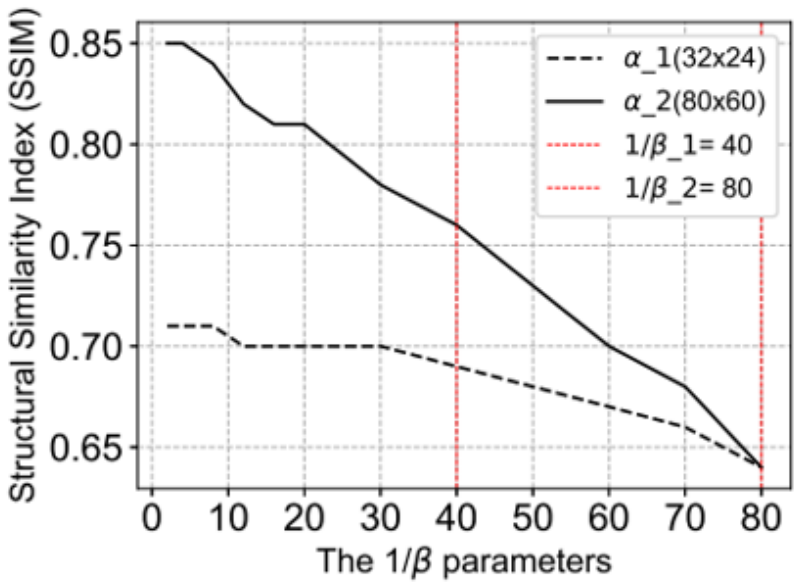
- The proof of an approximately Gaussian.



Appendix - the alpha / beta parameters

- The α and β parameters are selected based on the results.





Appendix - the pairs of copies

- A common measure to characterized the PVCD dataset: [Jiang, 2014; He, 2022]
- A pair of copy is a combination of two positive videos.
- A total number of pairs of copies are computed:

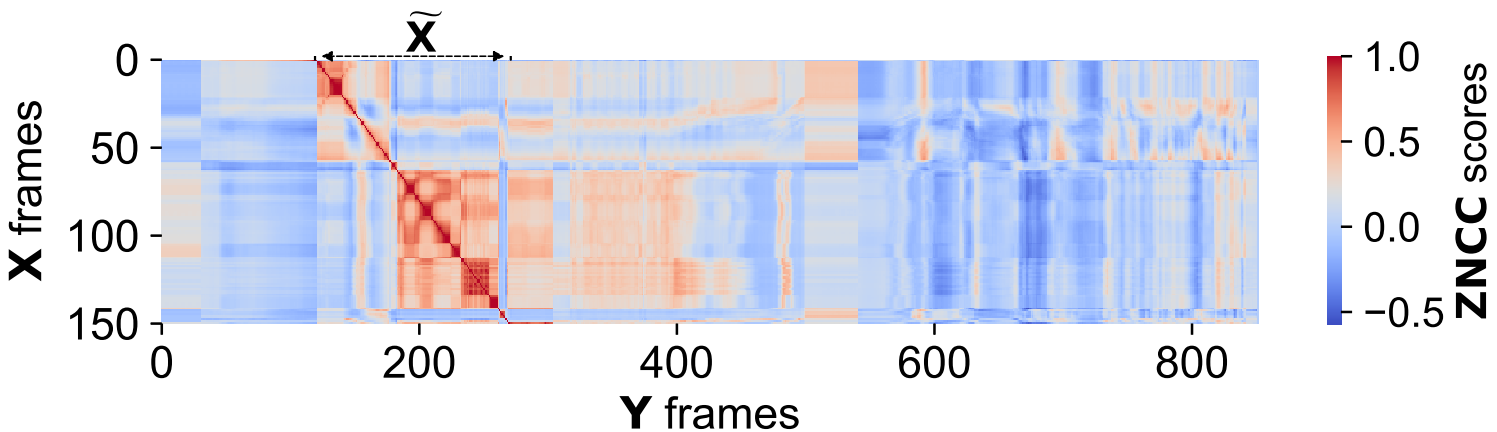
$\mathbf{C} = k \times \binom{n}{2}$, where k: #reference videos, n: #positive videos.

Appendix - GPU implementation (1/2)

- **ZNCC** between 2 images $\mathbf{I}, \mathbf{I}^*$ with means $\bar{\mathbf{I}}, \bar{\mathbf{I}}^*$ and standard deviations $\sigma_{\mathbf{I}}, \sigma_{\mathbf{I}^*}$ are given as follows.

$$\mathbf{ZNCC}(\mathbf{I}, \mathbf{I}^*) = \frac{\sum_{\forall \mathbf{x}} (\mathbf{I}(\mathbf{x}) - \bar{\mathbf{I}})(\mathbf{I}^*(\mathbf{x}) - \bar{\mathbf{I}}^*)}{\sigma_{\mathbf{I}} \sigma_{\mathbf{I}^*}}$$

- A visual frame-matching between two videos, X, Y.



Appendix - GPU implementation (2/2)

- GPU implementation: 5 steps
  1. downscaling and computing zero-mean images for $\mathbf{X}$,
  2. off-line computing the σ for $\mathbf{X}$,
  3. applying steps 1, 2 to the $\mathbf{Y}$,
  4. calculating a global matching of matrix $\mathbf{S}$,
  5. calculating the vector result $\mathbf{W}$ with a weighted method.

$$\mathbf{S} = \begin{bmatrix} \mathbf{s_{00}} & \cdots & \mathbf{s_{0n}} \\ \vdots & \mathbf{s_{kl}} & \vdots \\ \mathbf{s_{m0}} & \cdots & \mathbf{s_{mn}} \end{bmatrix}, \text{ where } \mathbf{w_l} = \sum_{\mathbf{k=0}}^{\mathbf{m}} \mathbf{s_{k,l+k}} \times \lambda_{\mathbf{k}}$$

- Time-efficient processing with GPU.

	CPU threads		GPU
	1	14	
Time processing (s)	420	390	5

Gray images size: 64x48, reference, testing videos length: 3s, 1200s.

Appendix - the separability

- The computational complexity for a full-search strategy:
  - #Ref the set of reference / jingle having size **k**,
  - #Post he set of positive video having a size **m**,
  - #Neg the set of negative video having a size **n**,

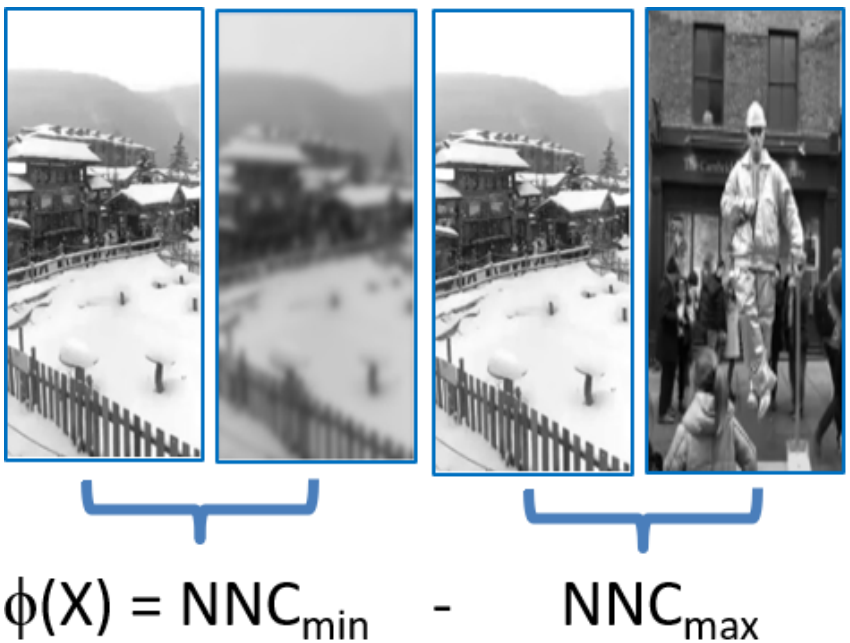
$$\mathbf{O}() = a + b + c, \text{ where } a = \sum_{i=1}^{k-1} i;\ b = k \times m;\ c = k \times n.$$

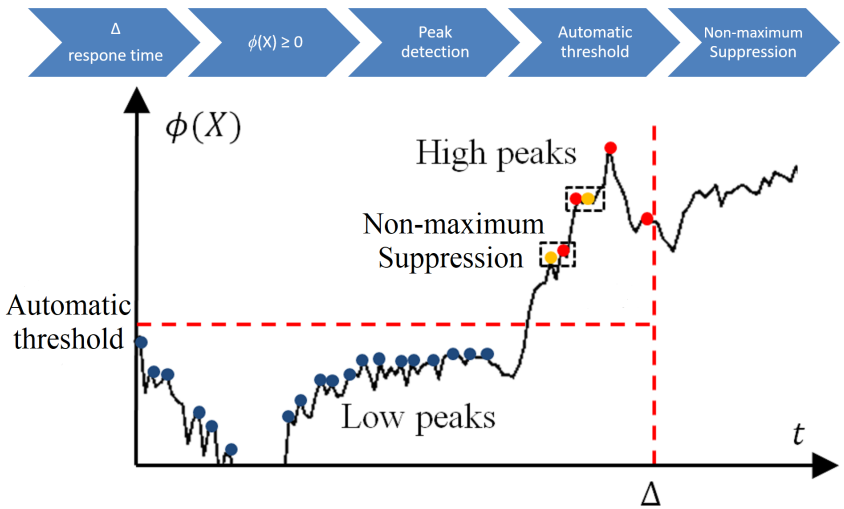
- The computational complexity for a comparison:
  - the average reference video having a duration: 3 s,
  - the window size having a duration: $20 \times 60 = 1200$ s,
  - the FPS of the video having: 30,

$$\mathbf{O}() = (3 \times 30) \times (1200 \times 30) = 3,240,000.$$

Appendix - key-frame selection (1/2)

- Real-time NCC: robustness, time optimization, predictability





Appendix - key-frame selection (2/2)

- The pipeline processed on the SVD dataset [Jiang, 2019].

1K videos
#Query set

27K videos
#Negative pairs

Aspect ratio normalization and gray-level conversion

(*A*, *B*) TN videos

(*C*) TP videos

Blurring, compression and contrast-alteration

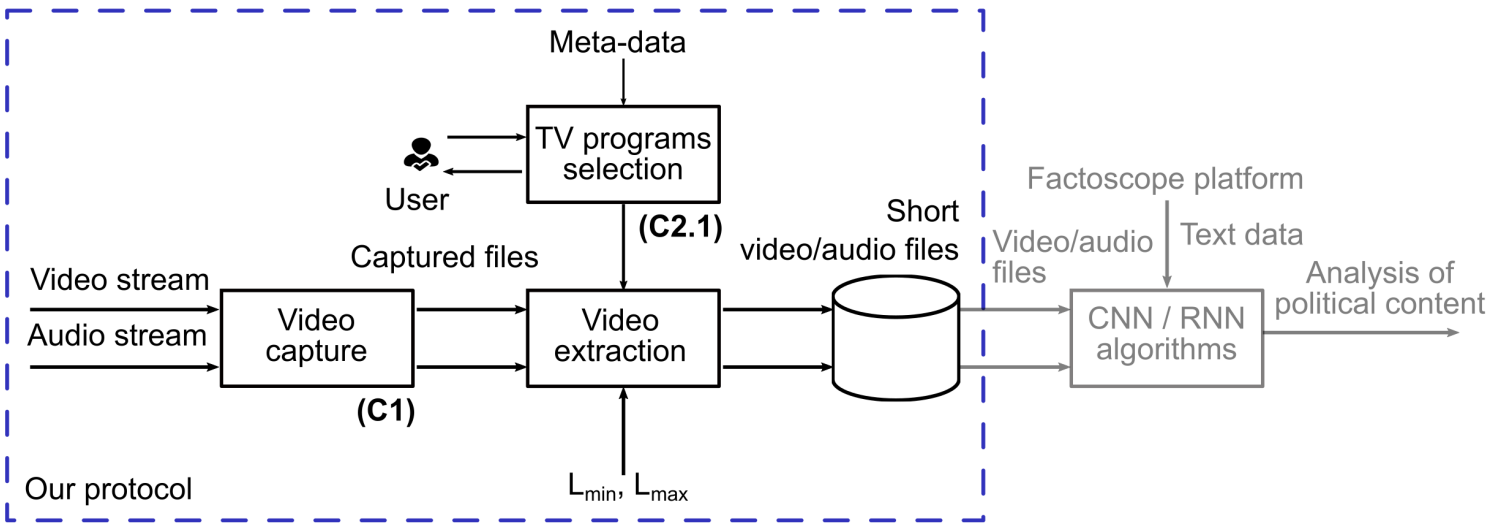
(*D*, *E*) TP videos

Training		*Testing*	
(*A*) TN	(*C*+*D*) TP	(*B*) TN	(*E*) TP
2K	1K+1K	25K	2K

K is thousand

Appendix - the fact checking dataset

- Protocol adaptation to design a dataset for fact-checking



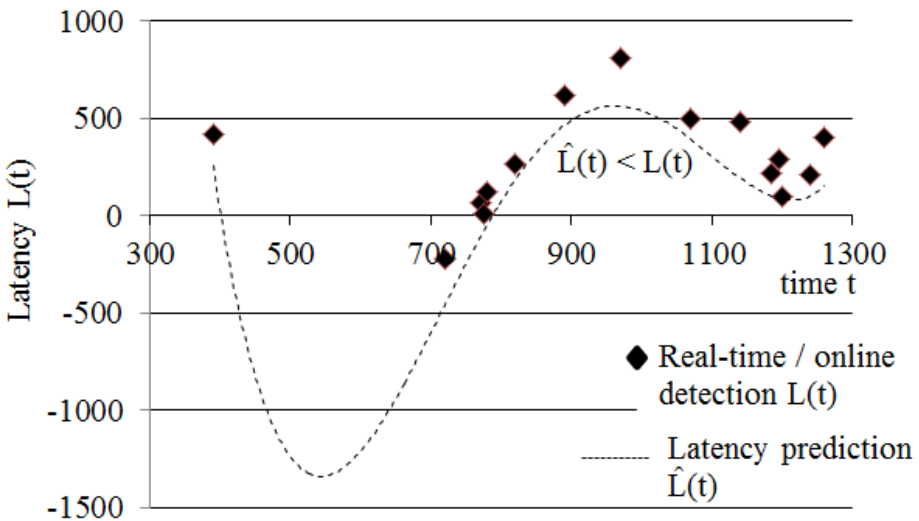
Captured total **(14,400 h)**	Meta-data programs	Main programs	Selected programs	Extracted programs
TV programs	16,943	14,172	6,293	
Hashcodes	998	245	93	
Duration (h)	9,918	7,933	4,105	5,826

Capture of 8 channels during period of the French presidential election (01/02/2022 up to 01/05/2022).

Appendix - the parallel machine scheduling

Problem statement: a full capture (32 channels) is costly (32k[7] + 5k / year for storage), a partial capture is required[8]. A dedicated PMS with help for optimization:

- no preemptive scheduling with static execution times of jobs, off-line optimal solution to refine online in real-time due to the latency $L(t)$,
- a public available dataset STVD-PMS[9] with an UT agreement (170 days, 26 channels, 99k jobs, 5,615 hashcodes, offline/online latency),
- PMS methods based meta-heuristic are investigated.



[7] Desktop version without maintenance and hosting cost. The currency unit is Euro.

[8] Not repeated / idle (40% × 70% = 28%), political, entertainment, etc.

[9] https://dataset-stvd.univ-tours.fr/pms/

References I


Chenot, J.H. et al (2014). "A large-scale audio and video fingerprints-generated database of tv repeated contents". In: *International Workshop on Content-Based Multimedia Indexing (CBMI)*.


Daghaghi, S. et al (2021). "Accelerating SLIDE Deep Learning on Modern CPUs: Vectorization, Quantizations, Memory Optimizations, and More". In: *Machine Learning and Systems (MLSys)*.


Guzman, Z.J. et al (2016). "A simple approach towards efficient partial-copy video detection". In: *International Workshop on Multimedia Signal Processing (MMSP)*.


Guzman, Z.Z.J et al (2019). "Partial-copy detection of non-simulated videos using learning at decision level". In: *Multimedia Tools and Applications*.


Han, Z. et al (2021). "Video similarity and alignment learning on partial video copy detection". In: *ACM International Conference on Multimedia*.


He, S. et al (2022). "A Large-scale Comprehensive Dataset and Copy-overlap Aware Evaluation Protocol for Segment-level Video Copy Detection". In: *Computer Vision and Pattern Recognition (CVPR)*.


Hou, Y.W. et al (2018). "Video copy detection based on uniform local binary pattern". In: *DEStech Transactions on Computer Science and Engineering*.


Hu, Y. et al (2019). "STRNN: End-to-end deep learning framework for video partial copy detection". In: *Journal of Physics: Conference Series*.

References II

Jiang, Q.Y. et al (2019). "SVD: A large-scale short video dataset for near-duplicate video retrieval". In: *International Conference on Computer Vision (ICCV)*.

Jiang, Y.G. et al (2014). "VCDB: a large-scale database for partial copy detection in videos". In: *European Conference on Computer Vision (ECCV)*.

Joly, A. et al (2007). "Content-based copy retrieval using distortion-based probabilistic similarity search". In: *Transactions on Multimedia*.

Kordopatis, Z.G. et al (2017). "Near-duplicate video retrieval by aggregating intermediate cnn layers". In: *International conference on Multimedia Modeling (MMM)*.

Law-To, J. et al (2007). "Video copy detection: a comparative study". In: *International Conference on Image and Video Retrieval (CIVR)*.

Liang, T. et al (2021). "Pruning and quantization for deep neural network acceleration: A survey". In: *Neurocomputing*.

Liu, X. et al (2021). "GANN: A Graph alignment neural network for video partial copy detection". In: *Conference on Big Data Security on Cloud (BigDataSecurity)*.

Özbulak, G. et al (2016). "Robust video copy detection in large-scale TV streams using local features and CFAR based threshold". In: *International Conference on Digital Signal Processing (DSP)*.

Tan, W. et al (2022). "A Fast Partial Video Copy Detection Using KNN and Global Feature Database". In: *Winter Conference on Applications of Computer Vision (WACV)*.

References III

Wu, X. et al (2007). "Practical elimination of near-duplicates from web video search". In: *ACM international conference on Multimedia (MM)*.

Zhang, C. et al (2020). "Large-scale video retrieval via deep local convolutional features". In: *Advances in Multimedia*.

Zhang, Y. et al (2016). "Effective real-scenario video copy detection". In: *International Conference on Pattern Recognition (ICPR)*.

Zhu, Y. et al (2016). "Large-scale video copy retrieval with temporal-concentration sift". In: *Neurocomputing*.