

LIFAT laboratory
University of Tours

Performance Evaluation for Scalable Partial Video Copy Detection

October 07th, 2024

Presented by LE Van Hao

Jury Members

DUFFNER Stefan	:	Maître de Conférences HDR, Université de Lyon, France	: Reviewer
GROSSI Giuliano	:	Associate Professor, Université de Milan, Italie	: Reviewer
CARDOT Hubert	:	Professeur, Université de Tours, France	: PhD director
DELALANDRE Mathieu	:	Maître de Conférences, Université de Tours, France	: Co-supervisor
DOMENGER Jean-Philippe	:	Professeur, Université de Bordeaux, France	: Examiner
BURIE Jean-Christophe	:	Professeur, Université de La Rochelle, France	: Examiner
PHAM The-Anh	:	Associate Professor, Université de Hong Duc, Vietnam	: Invited

Outline

- ① Introduction to partial video copy detection (PVCD)
- ② Part I: A large-scale TV dataset for PVCD
- ③ Part II: Performance evaluation for scalable PVCD
- ④ Conclusions & perspectives

Introduction to partial video copy detection (PVCD)

Introduction (1/3)

- ▶ Partial video copy detection (PVCD) aims at finding one/more short video segments which have been transformed into a longer video.

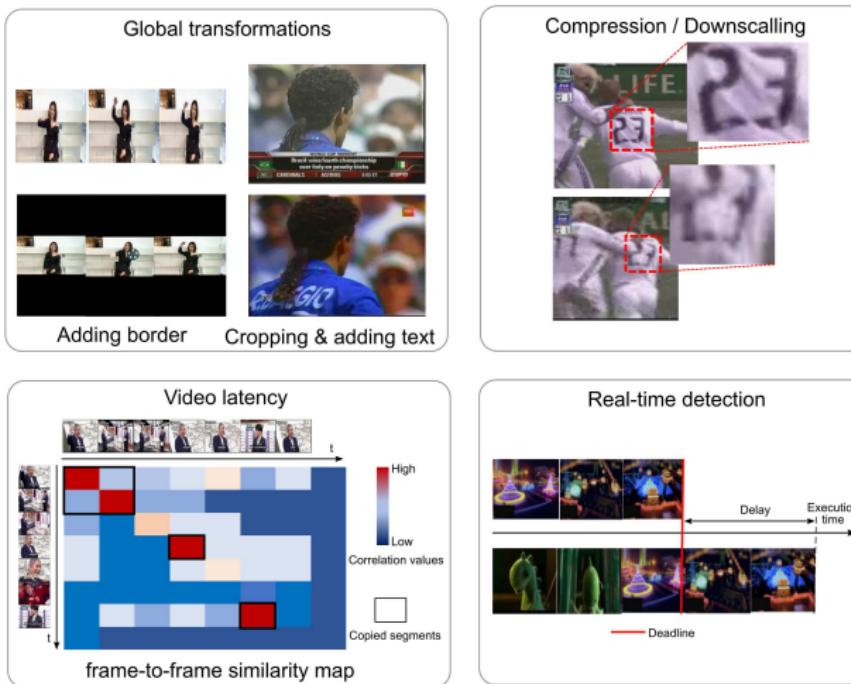


- ▶ Applications: copyright protection, video retrieval & monitoring, etc.
- ▶ A key topic in Computer Vision [Guzman, 2019; Liu, 2021; Tan, 2022],
Meta AI contest 2023¹

¹CVPR: <https://sites.google.com/view/vcdw2023/video-similarity-challenge>

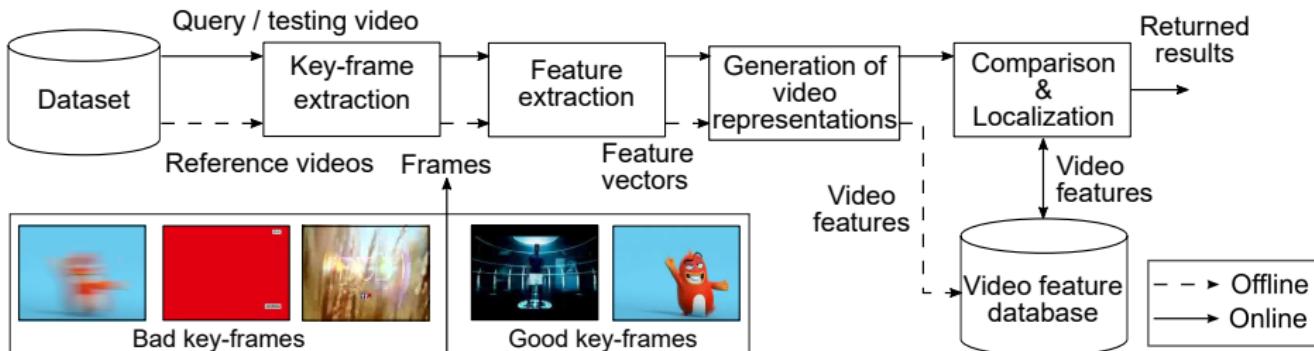
Introduction (2/3)

- ▶ Diversity in detection problems [Kordopatis, 2017; Liu, 2021; Tan, 2022]



Introduction (3/3)

- ▶ Common components of a typical PVCD system [Jiang, 2016; Hu, 2019]



- ▶ Top competitive systems use 2D CNNs [Han, 2021; He, 2022].
- ▶ Little discussion on improvement & characterization of these features [Kordopatis-Zilos, 2017; Hu, 2019]
- ▶ Dataset's limitations² avoid fine characterization/training [He, 2022]

²low scalability, imbalanced, temporal accuracy, uncontrolled noise

Part I: A large-scale TV dataset for PVCD

Part I: state-of-the-art PVCD datasets

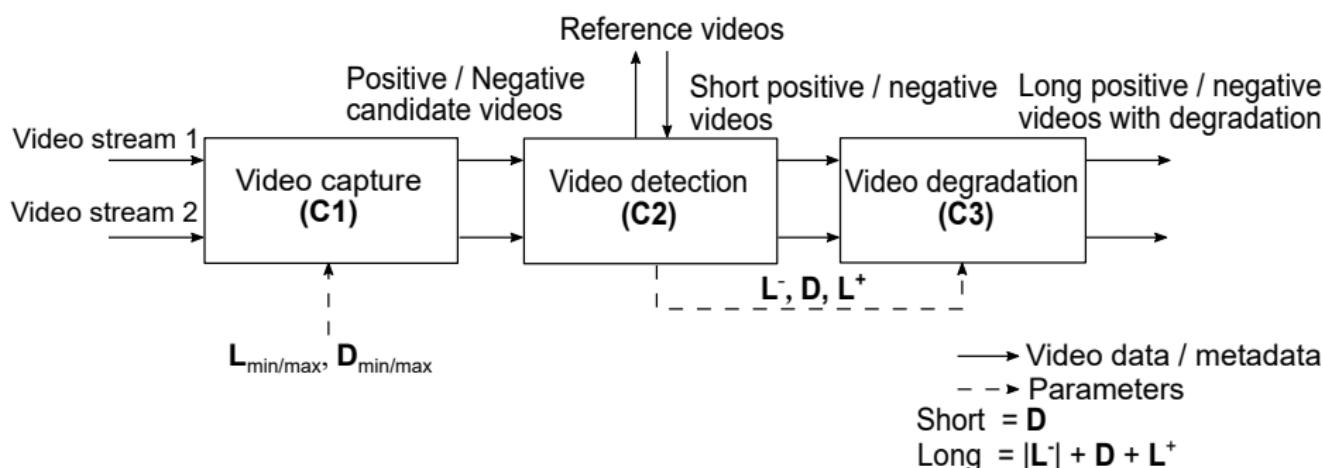
- ▶ 3 relevant datasets: VCDB, FIVR-PVCD, VCSL
- ▶ Protocols: Web videos with real degradations, manual annotation

Dataset	VCDB [Jiang, 2014]	FIVR-PVCD [Han, 2021]	VCSL [He, 2022]
References	small	small	small
Positive pairs	small	middle	large
Annotation cost	+	++	+++
Negative videos	middle	no	no
Overall scale	middle	middle	large
Degradation	real	real	real
Annotation level	segment	segment	segment
Timestamps (s)	1	1	1

- ▶ None of them fully reflect the real-world complexities of PVCD.

Part I: Our protocol (1/7)

- ▶ We propose a new protocol extended from [Joly, 2007; Law-To, 2007].
- ▶ 3 main components of our protocol:
 - ▶ **(C1)** captures videos with a TV workstation,
 - ▶ **(C2)** annotates reference, positive, and negative videos,
 - ▶ **(C3)** generates test sets with synthetic degradations.



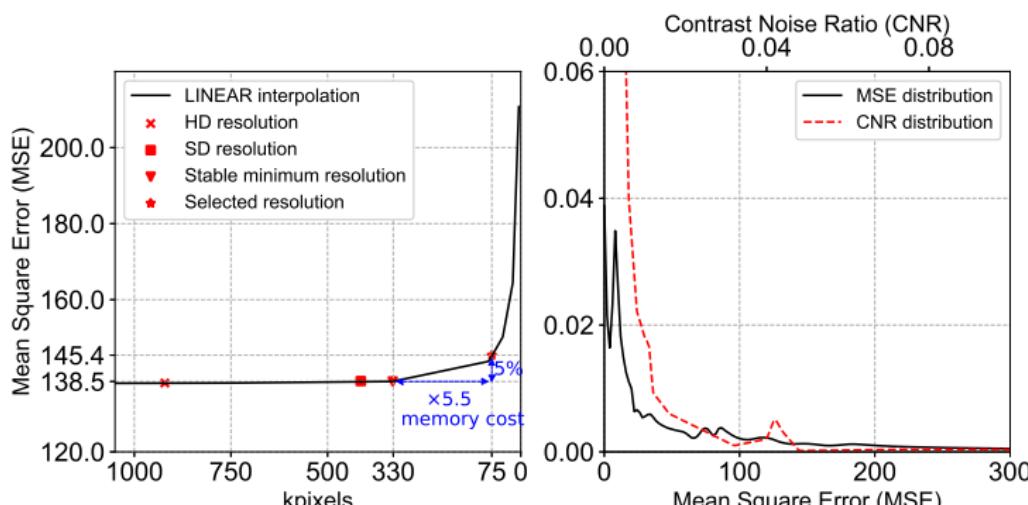
Part I: Our protocol (2/7)

Video capture (C1)

- ▶ It captures French TVs (MPEG-4/H264) using a TV workstation.

Channels per month	Daily file					Full capture		
	Resolution	kbps	Aspect ratio	Length	FPS	Months	Files	Size
8	320 × 240	560	4 : 3	20 h	30	3	720	3.46 TB

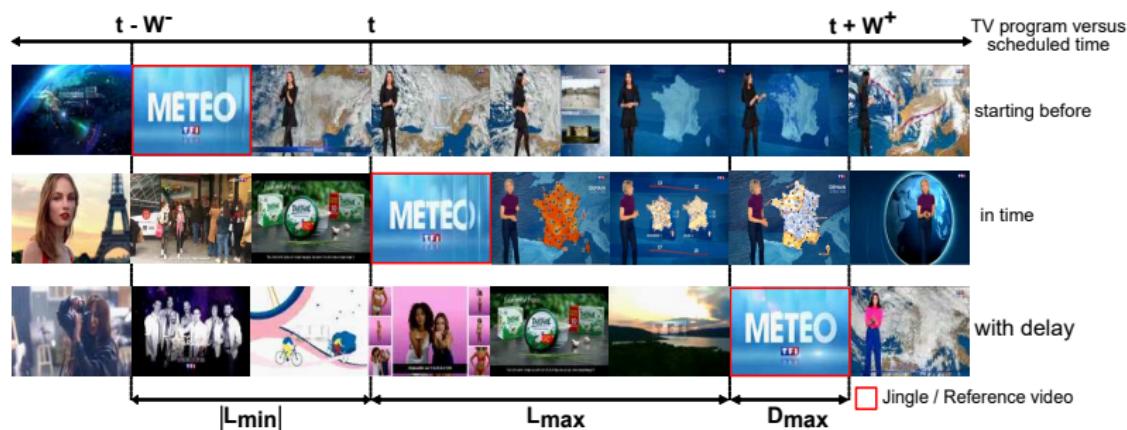
- ▶ An optimal strategy for memory cost and noise level



Part I: Our protocol (3/7)

Video capture (C1)

- ▶ Selection of positive candidates using metadata information:
 - ▶ collecting the TV metadata from EPG³,
 - ▶ selecting daily programs with high frequency,
 - ▶ hashing/encoding the names of these programs.
- ▶ The windows model to extract candidate & reference videos

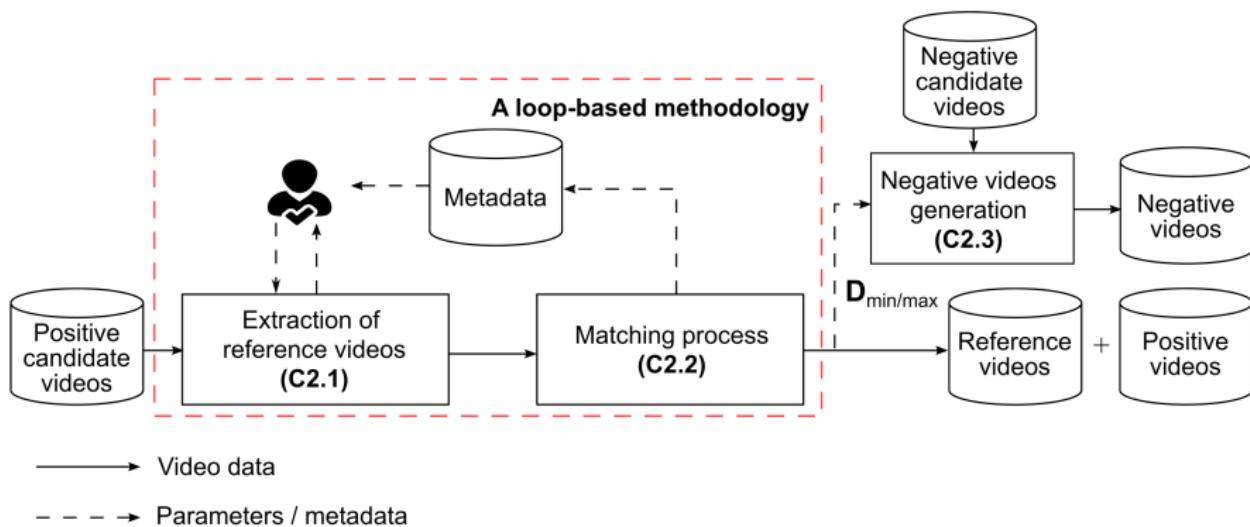


³EPG: Electronic Program Guide, e.g., <https://xmltv.ch>

Part I: Our protocol (4/7)

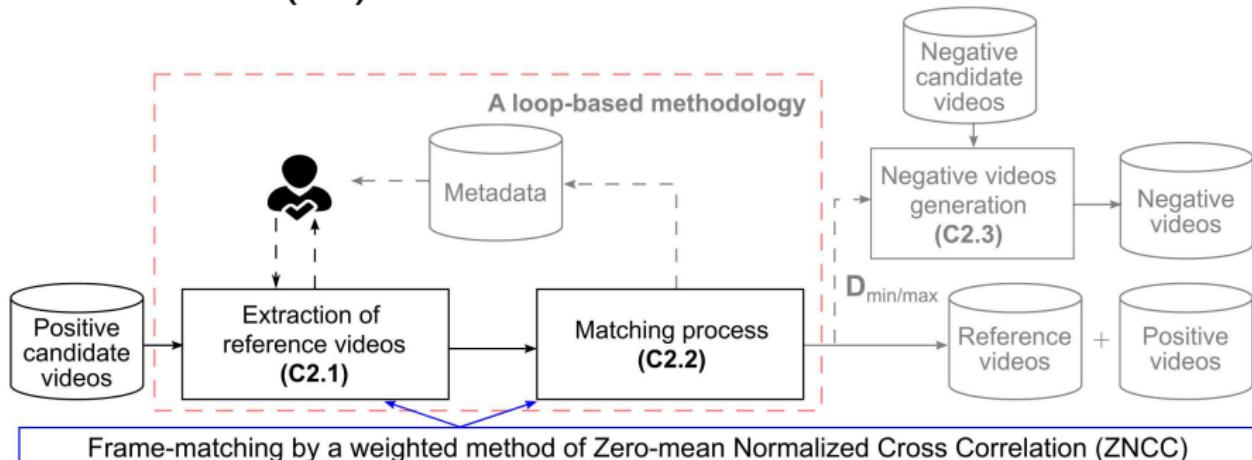
Video detection (C2)

- ▶ A semi-automatic annotation process with:
 - ▶ 3 sub-components **(C2.1), (C2.2), (C2.3)**,
 - ▶ a **loop-based methodology** to generate reference, positive videos,
 - ▶ an experiment-based procedure to generate negative videos.

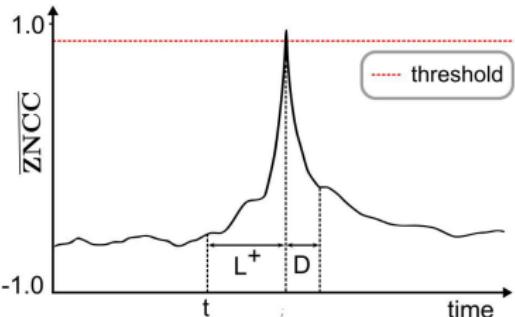
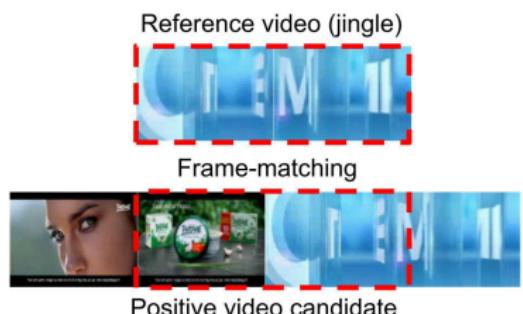


Part I: Our protocol (5/7)

Video detection (C2)

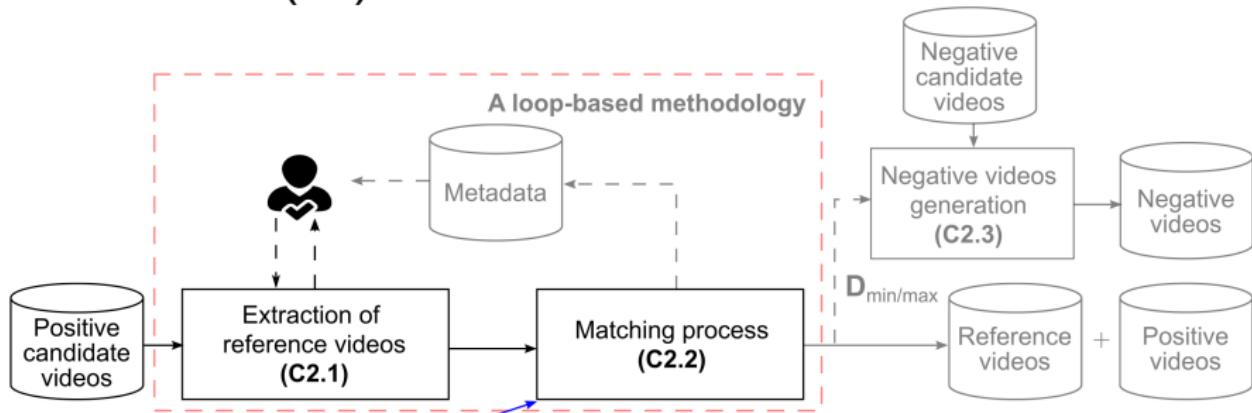


Frame-matching by a weighted method of Zero-mean Normalized Cross Correlation (ZNCC)

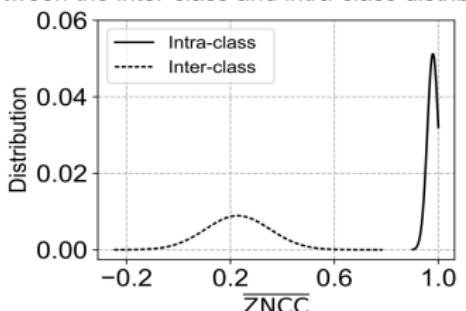


Part I: Our protocol (5/7)

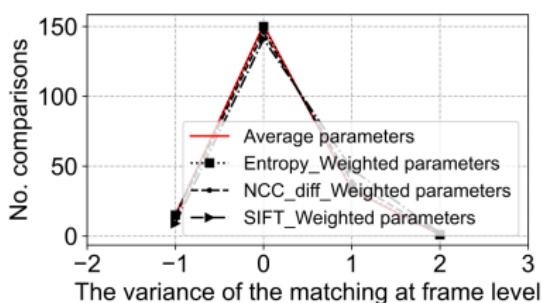
Video detection (C2)



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

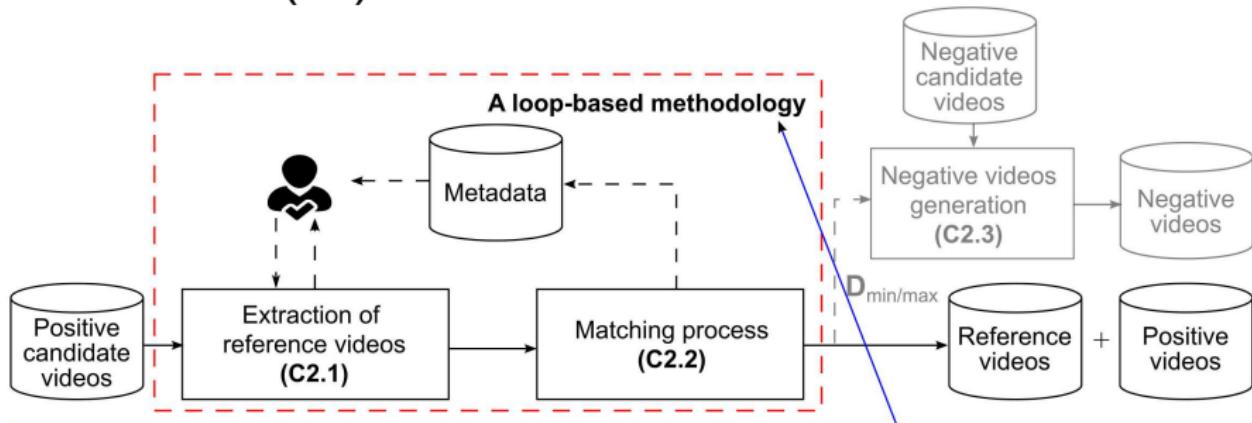


The accuracy of weighting methods for timestamping



Part I: Our protocol (5/7)

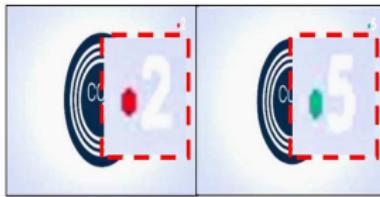
Video detection (C2)



A loop-based methodology to select/correct the reference videos to label while minimizing the user interaction cost.



different visual content



near-duplicate cases with logo change / text variation

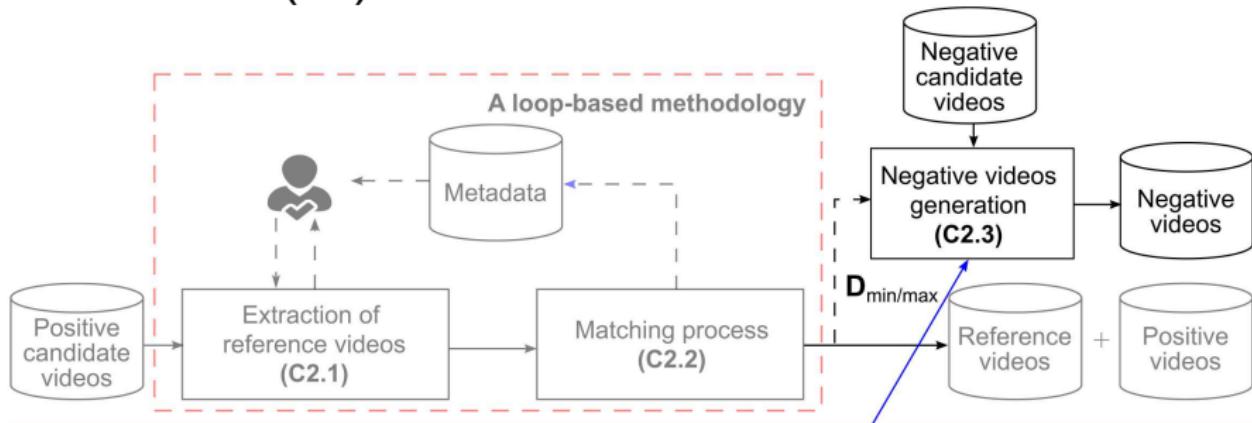
create a new reference sample



keep only one reference sample

Part I: Our protocol (5/7)

Video detection (C2)



Generate negative videos based on the distribution of the duration of reference videos

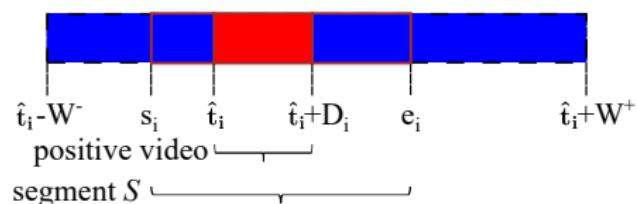
$$D \in [D_{\min}, D_{\max}]$$



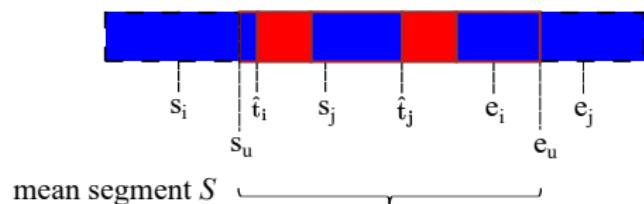
Part I: Our protocol (6/7)

Video degradation (C3)

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



(a)



(b)

Positive videos Negative videos Selected videos Not selected videos

Part I: Our protocol (7/7)

Video degradation (C3)

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

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



Part I: Experiments and results (1/2)

- ▶ 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

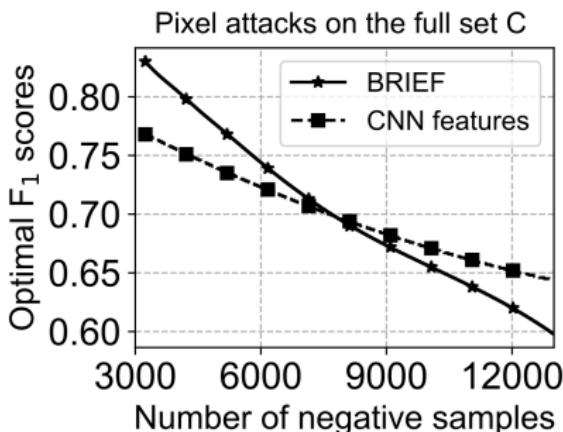
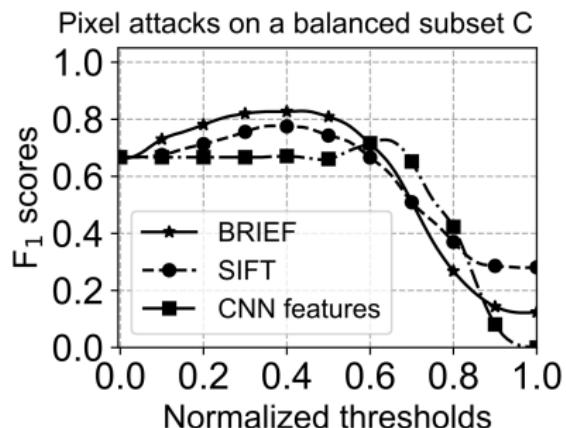
- ▶ The comparisons between STVD versus VCDB, VCSL

Datasets	VCDB [Jiang, 2014]	VCSL [He, 2022]	STVD Ours
Reference videos	28	122	243
Positive videos	528	9 207	19 280
Positive pairs	9 073	281 182	1 688 225
Negative videos	100 000	n/a	64 040
Duration (h)	2 000	n/a	10 660
Noise characterization	real noise	real noise	noise-free
Annotation cost (m-h)	700	20 000	105
Frame-level annotation	X	X	✓

(h): hours, (m-h): man-hours, and n/a: not available

Part I: Experiments and results (2/2)

- ▶ Baseline methods for evaluation [Zhu, 2016; Zhang, 2016; Zhang, 2020]
 - ▶ Protocol: key-frame extraction & matching⁴, segment-level / no-training
 - ▶ Metrics: Precision, Recall, F-measure
 - ▶ Test sets: B for a 'Hello world' ability $F_1 \simeq 0.98$, C for pixel attacks



- ▶ Local features are more effective, CNN features can deal with scalability.

⁴Correctly detected: at least one frame hits the groundtruth

Part I: Summary of results & perspectives

- ▶ We proposed a protocol to design a PVCD dataset:
 - ▶ ensuring the dataset scalability with balanced data,
 - ▶ offering a fine control of degradation,
 - ▶ and a null timing deformation with frame-level accuracy.
- ▶ We published the best PVCD dataset, named STVD that
 - ▶ includes $\simeq 83$ k videos, 10 k hours, 1688 k positive pairs,
 - ▶ evaluates baseline methods (hello-world/scalability testing),
 - ▶ is publicly available at <https://dataset-stvd.univ-tours.fr/>(PVCD),
 - ▶ has been promoted in the CV/ML research community⁵,
 - ▶ is designed with an adaptable protocol⁶.
- ▶ STVD dataset could be employed
 - ▶ for advanced performance evaluation of PVCD,
 - ▶ and to design new characterization methods.

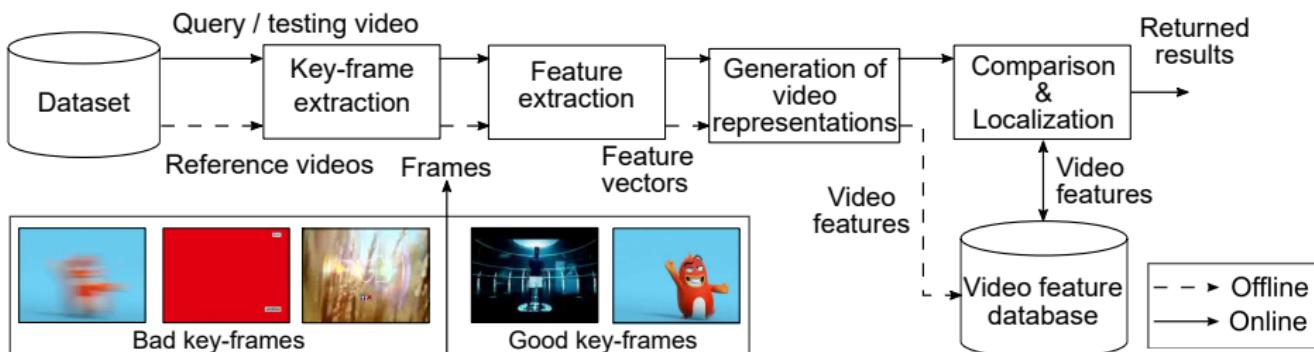
⁵cited papers: 3, CV/ML web platforms: 6, research mailings: 14, email contacts: 40.

⁶Fact checking [Rayar, 2022], operation research: <https://dataset-stvd.univ-tours.fr/pms/>

Part II: Performance evaluation for scalable PVCD

Part II: PVCD systems using 2D CNN features (1/2)

- ▶ Popularity 2D CNN features [Kordopatis, 2017; Han, 2021; Tan, 2022]

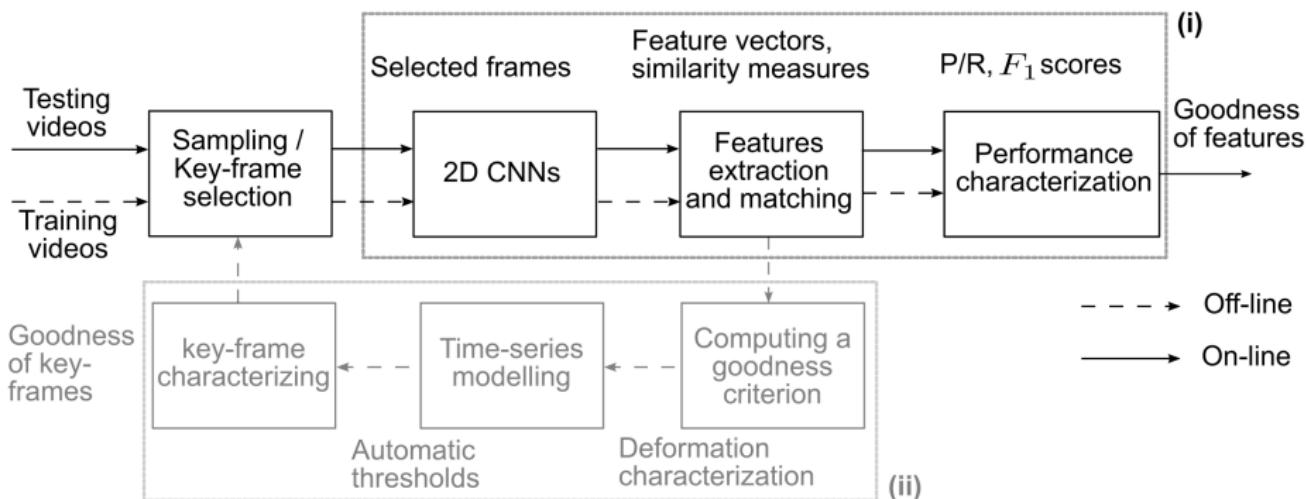


- ▶ Variety of 2D CNN features proposed [Phalke, 2020]
- ▶ 9 main pre-trained 2D CNN features are used for PVCD, based:
 - ▶ 3 CNN models (VGGNet, InceptionNet, ResNet),
 - ▶ 3 extraction techniques (Last FC, MAC, R-MAC)⁷.
- ▶ Their characterization has been little discussed [Hu, 2019].

⁷Last FC: obtained from Last Fully Connected layer, MAC/R-MAC: obtained from Convolutional Layers

Part II: Our methods (1/2)

- ▶ We propose a characterization work in two steps:
 - ▶ (i) to evaluate the performance of popular 2D CNN features,
 - ▶ (ii) to characterize the goodness of key-frames with a dedicated method.



Part II: Our methods (2/2)

- ▶ Protocol: detection based on key-frames⁸ [Jiang, 2016]
- ▶ Metrics: standard P/R, F_1
- ▶ Dataset: STVD⁹ [Le, 2022]
 - ▶ test set D (global transformations)
 - ▶ sampling with high recommended/scalability fitting the GPU constraint
 - ▶ data split 6÷4 for training, testing

Videos	No	Duration (h)	FPS	60% training	40% testing	Total
Negative videos	12 165	1 545	0.08	259 050 f	172 700 f	431 750 f
Positive videos	3 869	415				
Copied segments	4 436	7.5	1	16 200 f	10 800 f	27 000 f
				Total ¹⁰ :	458 750 f	

⁸sampling/labeling of key-frames, matching with Cosine similarity, correctly detected: at least one key-frame shared

⁹large-scale, balanced positive/negative distribution, accurate timestamping

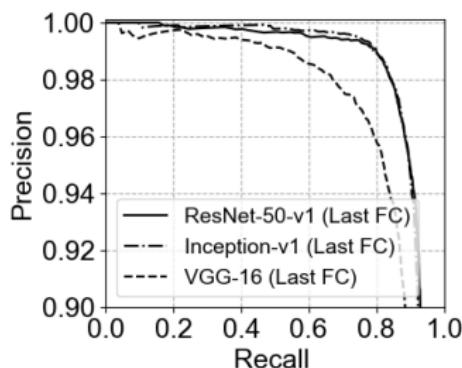
¹⁰Total: 458 750 f \simeq 7 GiB of 4 096-F dimension

Part II: Comparison results between CNN features

- ▶ Large-scale experiments to characterize these 2D CNN features:
 - ▶ 9 CNN features (3 models \times 3 methods¹¹),
 - ▶ 4.4×10^6 vectors, 445×10^9 matchings, optimized implementation.
- ▶ Comparison of 2D CNN feature results

Tab. 1 Top F_1 scores

	Last FC	MAC	R-MAC
ResNet50-v1	0.926	0.828	0.823
Inception-v1	0.923	0.738	0.782
VGG-16	0.894	0.922	0.918

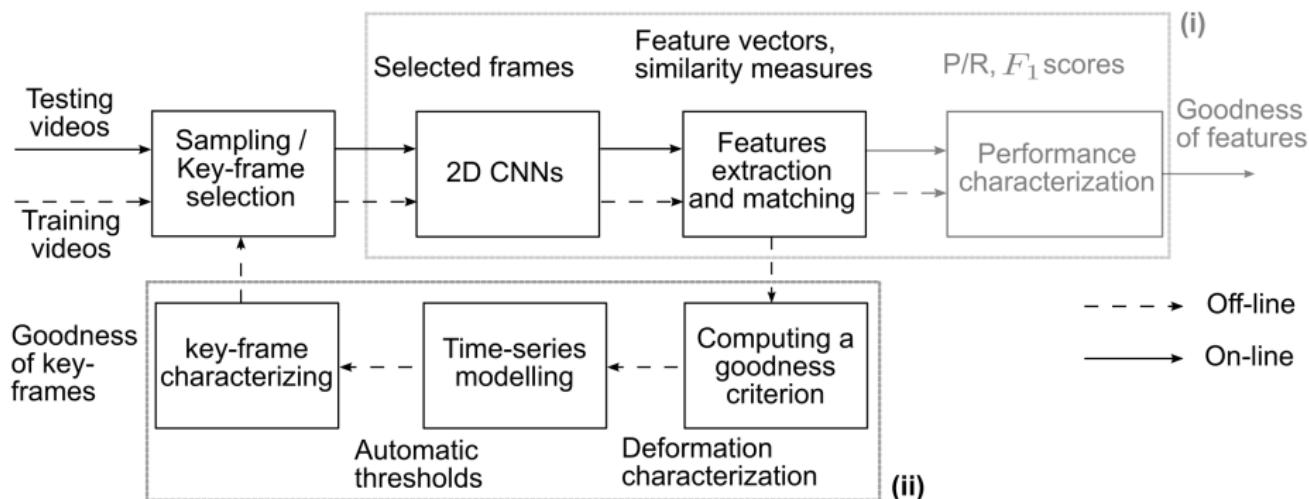


- ▶ Our results align with the state-of-the-art and highlight:
 - ▶ recent 2D CNN models (e.g., ResNet-50) outperform [He, 2016],
 - ▶ correlation between 2D CNN & methods (VGG & MAC) [Cools, 2022],
 - ▶ challenges remain in achieving full separability (even if $F_1 \simeq 0.93$).

¹¹Last Fully Connected (Last FC), Maximum Activations of Convolutions (MAC), and Regional-MAC (R-MAC)

Part II: Our methods to characterize key-frames (1/4)

- ▶ We propose a characterization work in two steps:
 - ▶ (i) to evaluate the performance of popular 2D CNN features,
 - ▶ (ii) to characterize the goodness of key-frames with a dedicated method.



Part II: Our methods to characterize key-frames (2/4)

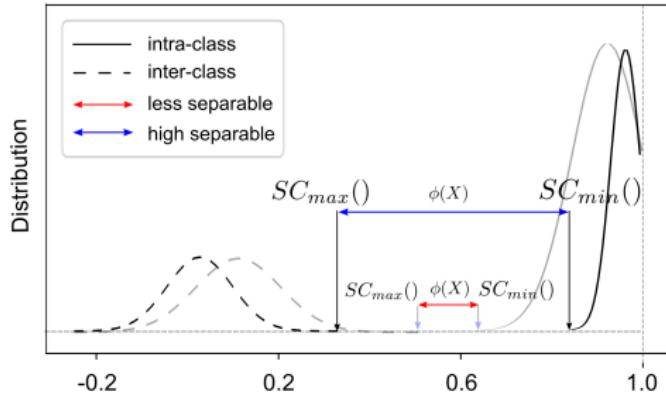
- ▶ We propose a goodness criterion $\phi(X)$ to evaluate the X separability
 - ▶ using Cosine similarity¹² ($SC \in [0, 1]$), then $\phi(X) \in [-1, 1]$ is defined as

X is a feature vector

$$\phi(X) = SC_{\min}(X, \{\tilde{X}_1, \dots, \tilde{X}_m\}) - SC_{\max}(X, \{Y_1, \dots, Y_{n_1}\}, \{X_1^*, \dots, X_{n_2}^*\})$$

↑ ↑
 \tilde{X} are near-duplicates of X Y are negatives,
 $X^* \neq X$ are different references

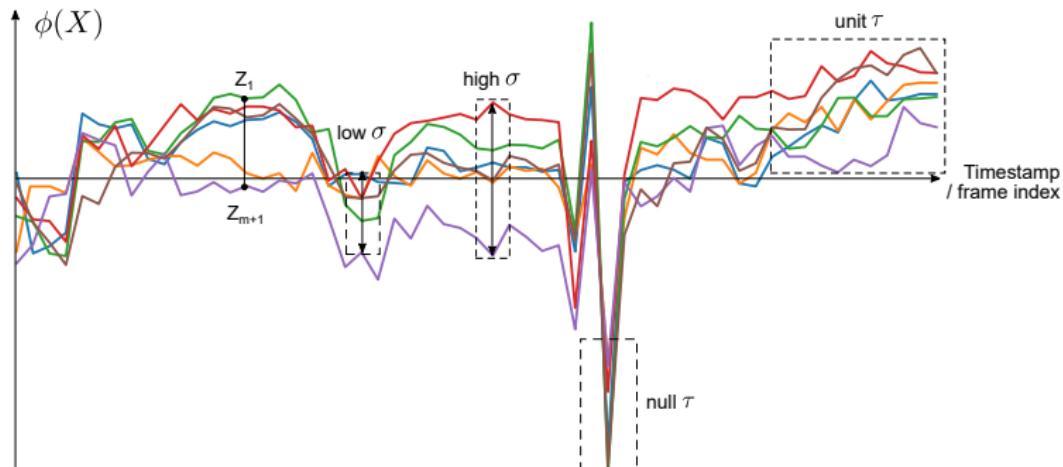
- ▶ Good key-frames support to maximize 'intra' & minimize 'inter'.



¹² $SC(X, Y) \in [0, 1]$ (non-negative vectors: ReLU, unit scale: L₂-norm) between vectors X, Y where $SC(X, X) \equiv 1$ / 38

Part II: Our methods to characterize key-frames (3/4)

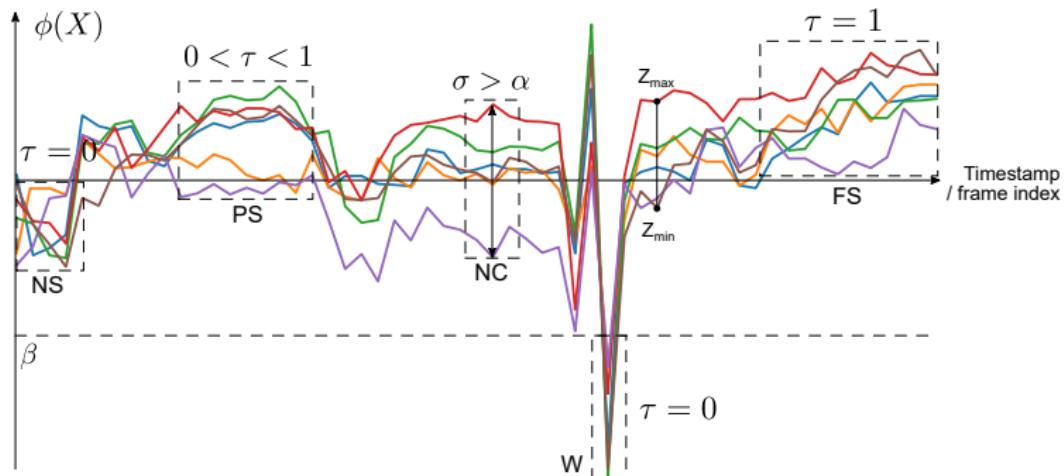
- ▶ The characterization of key-frames is done with time series:
 - ▶ modeling the criterion in the video collection and timing domains,
 - ▶ z_1, \dots, z_{m+1} values at t give the $\phi(X), \phi(\tilde{X}_1), \dots, \phi(\tilde{X}_m)$ criteria,
 - ▶ feature vector X and its near-duplicates $\{\tilde{X}_1, \dots, \tilde{X}_m\}$.



- ▶ Detect consistency & separability of key-frames from z_1, \dots, z_{m+1} :
 - ▶ the standard deviation σ ,
 - ▶ a rate τ accounting the amount of positive criteria.

Part II: Our methods to characterize key-frames (4/4)

- ▶ We compare σ , τ to automatic thresholds α , β for categorization:
 - ▶ $\sigma > \alpha$ serves to filter Not Consistent (NC),
 - ▶ $\tau \neq 0$ serves to get Partially Separable (PS), Fully Separable (FS),
 - ▶ β is an adaptive threshold to split Not separable (NS) / Worst (W).



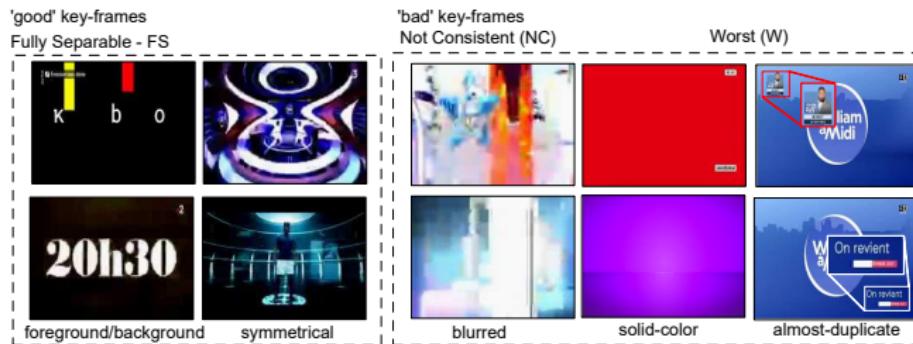
Part II: Our results & key-frame categories

- We extended (i) dataset for the feature VGG-16/MAC¹³ at 30 FPS.
(50×10^3 indices, 0.8×10^6 vectors, 244×10^9 matchings)

(NC-Not consistent, W-Worst, NS-Not separable, PS-Partially Separable, FS-Fully Separable.)

Total indices	NC	W	NS	PS	FS
50 844	6 966	4 169	33 049	4 881	1 780
100 %	13.7 %	8.2 %	65 %	9.6 %	3.5 %
	86.9 %			13.1 %	

- Key-frame categories:
 - $\simeq 4\%$ 'good' (FS) key-frames,
 - $\simeq 87\%$ 'hardly' separable (NC+W+NS), of which $\simeq 22\%$ 'bad' (NC, W)



¹³for a trade-off between a strong detection and the memory constraint ($F_1 = 0.922$ & 512-F dimension)

Part II: Summary of results

- ▶ We reported large-scale experiments to characterize 2D CNN features:
 - ▶ **9** CNN features, **4.4 m** vectors, **445 b** matchings,
 - ▶ ResNet-50 outperforms, correlation CNN & methods,
 - ▶ Scalability challenge offered by the STVD dataset.
- ▶ We proposed a method for the characterization of key-frames:
 - ▶ goodness criterion, time-series modelling, categorization method,
 - ▶ **50 k** indices, **0.8 m** vectors, **244 b** matchings,
 - ▶ a dominant amount of key-frames is not suitable for 2D CNN features.

Conclusions & perspectives

Conclusions (1/2)

- ▶ We proposed a protocol to design a PVCD dataset offering:
 - ▶ scalability, balanced data, fine control of degradation,
 - ▶ null timing deformation with frame-level accuracy.
- ▶ We published the best PVCD dataset, named STVD that:
 - ▶ includes $\simeq 83 \text{ k}$ videos, 10 k hours, 1688 k positive pairs, 5 test sets,
 - ▶ is publicly available and promoted in the CV/ML research community.

Conclusions (2/2)

- ▶ We conducted extensive experiments for performance evaluation.
 - ▶ Protocol: key-frame extraction and matching
 - ▶ Metrics: P , R , F_1
- ▶ We draw several key conclusions:
 - ▶ ability of our dataset for PVCD,
 - ▶ hand-crafted features process well on standard tasks,
 - ▶ characterization of PVCD 2D CNN features.
- ▶ We provided a fine characterization of key-frames:
 - ▶ goodness criterion, time-series modelling, and automatic thresholding,
 - ▶ a dominant amount of key-frames is not suitable for 2D CNN features.

Our perspectives

- ▶ Perspectives on our dataset:
 - ▶ promote (contest online / conference, communication),
 - ▶ generate new test sets for specific challenges,
 - ▶ provide more scalability¹⁴,
 - ▶ adapt our protocol to new research problems,
 - ▶ address a non-rigid PVCD problem.
- ▶ Perspectives on PVCD performance evaluation:
 - ▶ very-large scalability / brute-force (cluster of GPU),
 - ▶ re-training for CNNs [He, 2022],
 - ▶ other recent deep learning models (2D, 3D),
 - ▶ key-frame classification / selection for video representations.

¹⁴Two strategies for collaborative and user interaction optimization / fully automatic annotation.

List of publications

- ① **V.H. Le**, M. Delalandre and H. Cardot, *Performance Characterization of 2D CNN Features for Partial Video Copy Detection*, Conference on Computer Analysis of Images and Patterns (CAIP), pp. 205-215, 2023.
- ② 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), pp. 181-185, 2022.
- ③ **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.
- ④ **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.
- ⑤ **V.H. Le**, M. Delalandre and D. Conte. *Real-time detection of partial video copy on TV workstation*. (CBMI), pp. 1-4, 2021.

Thank you for your attention !

References |

- Cools, A., M.A. Belarbi, and S.A. Mahmoudi (2022). "A Comparative Study of Reduction Methods Applied on a Convolutional Neural Network". In: *Electronics* 11, p. 1422.
- 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, K. (2016). "Deep residual learning for image recognition". In: *Conference on computer vision and pattern recognition (CVPR)*, pp. 770–778.
- 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)*.
- Hu, Y. et al (2019). "STRNN: End-to-end deep learning framework for video partial copy detection". In: *Journal of Physics: Conference Series*.
- Jiang, Y.G. and J. Wang (2016). "Partial copy detection in videos: A benchmark and an evaluation of popular methods". In: *IEEE Transactions on Big Data* 2.1, pp. 32–42.
- 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*.

References II

- Kordopatis, Z.G. et al (2017). "Near-duplicate video retrieval by aggregating intermediate cnn layers". In: *International conference on Multimedia Modeling (MMM)*.
- Kordopatis-Zilos, G. (2017). "Near-duplicate video retrieval with deep metric learning". In: *International Conference on Computer Vision Workshops (ICCV)*, pp. 347–356.
- Law-To, J. et al (2007). "Video copy detection: a comparative study". In: *International Conference on Image and Video Retrieval (CIVR)*.
- Le, V.H., M. Delalandre, and D. Conte (2022). "A large-Scale TV Dataset for partial video copy detection". In: *International Conference on Image Analysis and Processing (ICIAP)*. Vol. 13233. Lecture Notes in Computer Science (LNCS), pp. 388–399.
- 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)*.
- Phalke, D. A. and S. Jahirabadkar (2020). "A survey on near duplicate video retrieval using deep learning techniques and framework". In: *Pune Section International Conference (PuneCon)*. IEEE, pp. 124–128.
- Rayar, F., M. Delalandre, and V.H. Le (2022). "A large-scale TV video and metadata database for French political content analysis and fact-checking". In: *International Conference on Content-based Multimedia Indexing (CBMI)*, pp. 181–185.
- Shen, L. et al (2020). "Advance on large scale near-duplicate video retrieval". In: *Frontiers of Computer Science (Front. Comput. Sci.)* 14.5, pp. 1–24.

References III

- 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)*.
- Tolias, G., R. Sicre, and H. Jégou (2016). "Particular Object Retrieval With Integral Max-Pooling of CNN Activations". In: *International Conference on Learning Representations (ICLR)*, pp. 1–12.
- 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*.