Performance characterization of 2D CNN features for partial video copy detection

Van-Hao LE, Mathieu Delalandre, Hubert Cardot
firstname.lastname@univ-tours.fr

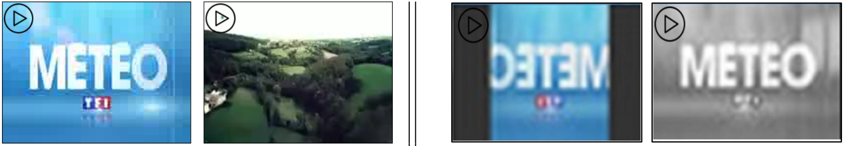
LIFAT Laboratory, University of Tours, France

September 26^{th}, 2023

Outline

Introduction

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

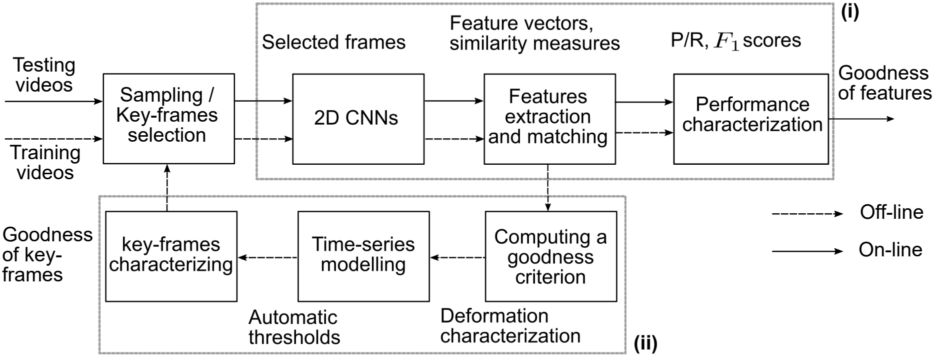


short segments || long videos which have transformed

- PVCD includes several application domains (copyright protection, video retrieval, etc.) [Han, 2021; Jiang, 2021; Tan, 2022].
- PVCD uses 2D CNN features, their characterization is little discussed for the PVCD task [Kordopatis-Zilos, 2017; Hu, 2019].

Objectives of our work

- **(i)** We report large-scale experiments to compare 2D CNN features:
  - comparison of 9 types of features with standard P/R and F_1 scores,
  - our conclusions and results are consistent with the CV state-of-the-art.
- **(ii)** We propose a method to characterize the goodness of key-frames:
  - a goodness criterion, time-series modelling & key-frames characterization,
  - highlights the difficulties of 2D CNN features for specific degradations.



Our work (1/2)

- **Video datasets**: VCDB [Jiang, 2016], SVD [Jiang, 2019], VCSL [He, 2022], and STVD [Le, 2022] which was selected[1].

Tab. 1 STVD Dataset (The h and s stand for in hours and in seconds).

Datasets	Degradation	Duration	References	Positive pairs	Timestamps
STVD	synthetic	10,660 h	243	1,688 K	$\frac{1}{30}$ s

Tab. 2 Pre-processing of the STVD dataset for our experiments.

Videos	60% training	40% testing	Total frames	Total frames
Negative videos	259,050 f	172,700 f	431,750 f	259,050 f
Copied segments	16,200 f	10,800 f	27,000 f	486,000 f
	(i)			**(ii)**

- **Protocols**: **(i)** standard P/R, F_1 scores using the Cosine similarity, **(ii)** a proposed protocol .

[1] fine control of degradations, large-scale, balanced positive / negative distribution, accurate timestamping

Our work (2/2)

- **Goodness criterion**: $\phi(X) \geq 0$ using Cosine similarity (SC) is given

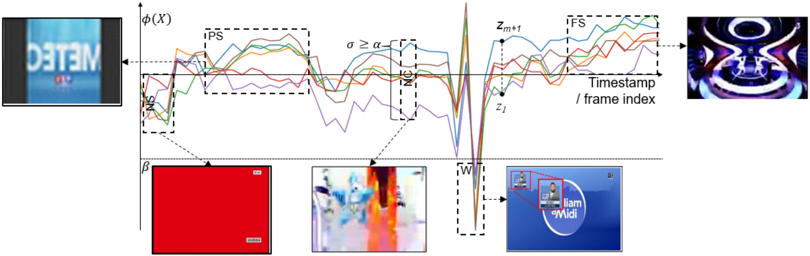
X is a feature vector

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

$\tilde{X}$ is near-duplicate of X

Y is negative,
$X^* \neq X$ has a different reference

- **Time-series modelling**:



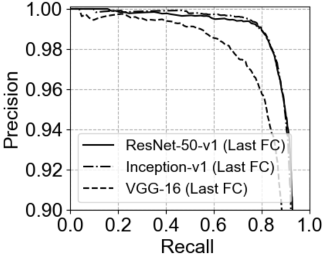
- **Key-frame categorization**: in 5 categories based on automatic thresholds, Not Consistent (NC), Worst (W), Not Separable (NS), Partially Separable (PS), Fully Separable (FS).

Main results (1/2)

- Large-scale experiments to characterize these 2D CNN features:
  - 9 CNN features (3 models × 3 methods[2]),
  - 4.4 **M** vectors, 445 **B** matchings.
- Comparison of 2D CNN features results.

Tab. 3 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 highlight and be consistent with the state-of-the-art:
  - the separability of features is not achieved (even if $F_1 \simeq 0.93$),
  - recent 2D CNN (ResNet-50) outperform [He, 2016],
  - correlation between 2D CNN & methods (VGG & MAC) [Cools, 2022].

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

Main results (2/2)

- A proposed method to characterize key-frames using 2D CNN features:
  - a goodness criterion, time-series modelling & key-frames categorization,
  - $\simeq 0.8$ **M** feature vectors, $\simeq 244$ **B** matchings.
- Results of key-frames categorization.

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

Total	NC	W	NS	PS	FS
100 %	13.7 %	8.2 %	65 %	9.6 %	3.5 %

- Our results highlight:
  - an 'easy' categorization of key-frames,
  - a quantitative analysis of the goodness of key-frames,
  - only a small amount of 'good' key-frames ($\simeq$ 13% in PS, FS),
  - difficulties to detect 'bad' key-frames ($\simeq$ 22% in NC, W).

Conclusions & Perspectives

- Our contributions for performance characterization of 2D CNN features
  - We report large-scale experiments to characterize 2D CNN features:
    - 9 CNN features, 4.4 M vectors, 445 B matchings,
    - ResNet-50 outperforms, correlation CNN & methods.
  - We propose a method for the characterization of key-frames:
    - goodness criterion, time-series, categorization,
    - 0.8 M vectors, 244 B matchings,
    - categorization and analysis, performance limits of features.
- Our perspectives to further improve the PVCD performance:
  - protocol of automatic labeling for scalable frame classification,
  - robust key-frame selection and learning of 2D CNN features.

Thank you for your attention!

References

Cheng, H., P. Wang, and C. Qi (2021). "CNN features based unsupervised metric learning for near-duplicate video retrieval". In: *Open-access repository (arXiv)*. 2105.14566v1.

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.

Han, Z. (2021). "Video similarity and alignment learning on partial video copy detection". In: *ACM International Conference on Multimedia (MM)*, pp. 4165–4173.

He, K. (2016). "Deep residual learning for image recognition". In: *Conference on computer vision and pattern recognition (CVPR)*, pp. 770–778.

He, S. (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)*, pp. 21086–21095.

He, S. (2023). "TransVCL: Attention-enhanced Video Copy Localization Network with Flexible Supervision". In: *AAAI Conference on Artificial Intelligence (AAAI)*.

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

Jiang, C. (2021). "Learning segment similarity and alignment in large-scale content based video retrieval". In: *ACM International Conference on Multimedia (MM)*, pp. 1618–1626.

Jiang, Q.Y. (2019). "SVD: A large-scale short video dataset for near-duplicate video retrieval". In: *International Conference on Computer Vision (ICCV)*, pp. 5281–5289.

References

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.

Kordopatis-Zilos, G. (2017). "Near-duplicate video retrieval with deep metric learning". In: *International Conference on Computer Vision Workshops (ICCV)*, pp. 347–356.

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.

Tan, W., H. Guo, and R. Liu (2022). "A fast partial video copy detection using KNN and global feature database". In: *Winter Conference on Applications of Computer Vision (WACV)*, pp. 2191–2199.