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

Van-Hao LE, Mathieu Delalandre, and Hubert Cardot

LIFAT Laboratory, RFAI group, Tours city, France
firstname.lastname@univ-tours.fr

Abstract. 2D CNN are main components for Partial Video Copy Detection (PVCD). 2D CNN features serve for the retrieval and matching of videos. Robustness is a key property of these features. It is a well-known problem in the computer vision field but little investigated for PVCD. The contributions of this paper are twofold: (**i**) based on a public video dataset, we provide large-scale experiments with 700 B of comparisons of 4.4 M feature vectors. We report conclusions for PVCD consistent with the state-of-the-art. (**ii**) the regular protocol for performance characterization is misleading for PVCD as it is bounded to the video level. A method for the characterization of key-frames with 2D CNN features is proposed. It is based on a goodness criterion and a time series modelling. It provides a fine categorization of key-frames and allows a deeper characterization of a PVCD problem with 2D CNN features.

Keywords: detection · video copy · 2D CNN · characterization

1 Introduction

Partial Video Copy Detection (PVCD) finds segments of a reference video which have transformed copies. It is a well-known topic in the computer vision field [10,21]. 2D CNN are main components to design PVCD systems. The systems extract 2D CNN features from frames for the retrieval and matching of videos. The performance characterization of 2D CNN features is a known topic in the computer vision field. However, it has been little investigated for PVCD.

The contributions of this paper are twofold: (**i**) based on a public video dataset, we provide large-scale experiments with 700 B of comparisons of 4.4 M feature vectors. These experiments report conclusions on the particular PVCD problem consistent with the state-of-the-art of the computer vision field. (**ii**) the regular protocol for performance characterization is misleading for PVCD as it is bounded to the video level. For a deeper analysis, we propose a method for the characterization of key-frames. This method applies a goodness criterion and a time series modelling. It provides a fine categorization of key-frames and allows a deeper characterization of a PVCD problem.

Section 2 provides a state-of-the-art. Section 3 details our performance characterization work. Conclusions and perspectives are discussed in Section 4. Table 1 gives the main symbols and mathematical notations used in the paper.

Table 1: Main symbols and mathematical notations used in the paper

Symbols	**Meaning**
K, M, B, F, f	thousand 10^3, million 10^6, billion 10^9, float and frame / feature vector
x, y, z	scalar values
m, n or m_i, n_j	sizes of sets / vectors with $i, j = 1, 2, \ldots$
$X = [x_1, \ldots, x_n], Y$	X is the feature vector of positive frame ($x_1, \ldots, x_n$ the elements), Y is negative
$\tilde{X}, X^*$	$\tilde{X} \simeq X$ is the near duplicate of X, $X^* \neq X$ has a different reference
$\{X_1, \ldots, X_n\}$	set of feature vectors
$\|\|X\|\|$	l_2-norm of X with $\|\|X\|\| = \sqrt{\sum_{\forall i} x_i^2}$
$X \cdot Y$	dot product between X and Y with $X \cdot Y = \sum_{\forall i} x_i y_i$
$SC(X, Y)$	Cosine similarity $SC(X, Y) = X \cdot Y \in [-1, 1]$ with $\|\|X\|\| = \|\|Y\|\| = 1$
$F_1 = 2\frac{P \times R}{P + R}$	F_1 score with P the precision and R the recall
$\phi(X)$	$= SC_{\min}(X, \{\tilde{X}_1, \ldots, \tilde{X}_m\}) - SC_{\max}(X, \{Y_1, \ldots, Y_{n_1}\}, \{X_1^*, \ldots, X_{n_2}^*\})$ the goodness criterion characterizing the separability with X when $\phi(X) \geq 0$
t, $[z_1, \ldots, z_{m+1}]$	observation at t with $[z_1, z_2, \ldots, z_{m+1}]$ the $\phi(X)$, $\phi(\tilde{X}_1)$, $\ldots$, $\phi(\tilde{X}_m)$ criteria
$z_{\min}, \overline{z}, z_{\max}, \sigma, \tau$	statistics of $[z_1, \ldots, z_{m+1}]$, with the minimum $z_{\min}$, mean $\overline{z}$ and maximum $z_{\max}$ values, σ the standard deviation and τ the rate of positive values $z_k > 0$
α, β	thresholds for categorization of frames
$\overline{\mathcal{Z}}$	mean of indices with $\tau = 0$ and $\sigma \leq \alpha$ for a reference to fix the threshold $\beta = \overline{\mathcal{Z}}$

2 Related work

2D CNN process images into convolutional layers and classify them using fully connected layers. When applied to PVCD, a pipeline embedding the 2D CNN must be defined for video processing Table 2. A first step is to select key-frames with sampling at fixed FPS. Closed key-frames in the temporal domain have redundancy. Adaptive methods have been proposed for elimination of 2D CNN features by K-means clustering or ranked inter-frame distances [1,19].

Table 2: Overview of PVCD systems using 2D CNN

Key-frame selection	• Fixed FPS [4,6,7,9,10,11,12,13,15,17,18,21,22] • Adaptive methods [1,19]
2D CNN	• VGGNet [9,11,13,15,18,19,21,22] • ResNet [4,7,8,11,15] • InceptionNet [1,11,12,17] • AlexNet[1,10,12,21]
Feature extraction	• Fully connected layers [1,10,11,12,13,18,21] • Convolutional layers [1,4,7,8,11,12,15,17,19,22] • Low-dimensional [4,6,18,19] • RoI based features [8,22]
Video matching	• Frame matching [7,13,15] • Global matching [1,4,6,7,8,15,21]

Key-frames are then processed with pre-trained 2D CNN such as AlexNet, VGGNet (16 and 19), ResNet (50, 101 and 152) and InceptionNet. They process input square matrixes $\in [224; 299]$ in the RGB colour space. They have different architectures and are delivered into different versions (1 to 4).

PVCD systems extract features from 2D CNN. These features serve for the retrieval and matching of videos. The common approach is to extract the features from the full frames even if a RoI based extraction can be applied [8,22]. The features can be obtained from (i) the Fully Connected (FC) layers (ii) or the convolutional ones. In the case (i), the Last FC is commonly used for extraction.

With convolutional layers (ii), standard methods have been established (e.g. MAC and R-MAC[1] [16]) used in several PVCD systems [8,22].

The videos are then matched from 2D CNN features. A first approach is to detect the videos from the matching of individual frames [13,15]. The matching can be made global with a frame-to-frame similarity matrix [1,4,6,8,15]. In both cases, it is common to apply a l_2 normalization to the features [9,11,12,15] and to match with the cosine similarity or the Euclidean distance. Low-dimensional approximations can be obtained with pooling [19] or PCA [1,6,18].

Robustness of 2D CNN features is a key property for the PVCD systems. The performance characterization of 2D CNN features is a known topic in the computer vision field. As a general trend, features extracted from recent 2D CNN perform better [5]. The MAC and R-MAC feature extraction methods are more adapted to the networks having large sizes of convolution layers [2]. The impact of blurring noise has been characterized in [14]. The ability of 2D CNN features to characterize particular images is highlighted in [20].

To the best of our knowledge, comparisons of 2D CNN for PVCD have been addressed only in [11,12,15,17]. The characterization has been done for global matching only. Datasets with a low-level of scalability (e.g. SVD [9]) [11,12,17] or unbalanced (VCDB [10]) [15,17] have been used. The fine characterization of 2D CNN features for PVCD has never been investigated.

3 Performance characterization of 2D CNN features

PVCD systems extract and match 2D CNN features. These features serve for the retrieval and matching of videos. Robustness is a key property of these features. It is a well-known topic in the computer vision field, however, it has been little investigated for PVCD. We provide in this section large-scale experiments to address this problem. We will introduce the video dataset and performance characterization protocol. Performance characterization results are discussed and conclusions are compared to the state-of-the-art of the computer vision field. A method for characterization of key-frames is then proposed for a deeper analysis.

3.1 Dataset and characterization protocol

For performance characterization, a dataset must be selected. Several main PVCD datasets have been proposed, Table 3 gives a comparison. We have selected the STVD[2] dataset [13]. This dataset has several key properties **(i)** it is captured from TV and is almost noise-free allowing a fine control of degradations with synthetic methods **(ii)** it is the largest dataset of the literature with ten thousand hours of video, 243 references and $1,688$ thousand positive pairs[3] **(iii)** it offers a balance distribution between the negative and positive videos **(iv)** it is delivered with an accurate timestamping for video alignment.

[1] Maximum Activations of Convolutions (MAC) and Regional-MAC (R-MAC)
[2] http://mathieu.delalandre.free.fr/projects/stvd/pvcd/
[3] A positive pair (v_i, v_j) is a combination of two partial video copies v_i and v_j [7,10].

Table 3: Datasets for PVCD performance evaluation
The h, s and N/A stand for in hours, in seconds and not available.

Datasets	**VCDB**	**SVD**	**STVD**	**VCSL**
Paper	[10]	[9]	[13]	[7]
Degradation	real	synthetic	synthetic	real
Duration (h)	2,030 h	197 h	10,660 h	17,416 h
References	28	1,206	243	122
Positive pairs	9 K	N/A	1,688 K	281 K
Timestamps (s)	1 s	N/A	$\frac{1}{30}$ s	1 s

From the videos and groundtruth of the STVD dataset we have applied a pipeline[4] to extract $458,750$ frames Table 4. These frames have been sampled from negative videos and copied segments and split into a training and a testing set. We have processed these frames with the 2D CNN VGG-16, ResNet50-v1 and Inception-v1 for characterization. These networks are typical for PVCD Table 2. The three common methods Last FC, MAC and R-MAC have been used for extraction with a l_2 normalization resulting in 9 databases for a total of 4.1 M of feature vectors (of dimensions 512-F, 1,024-F, 2,048-F and 4,096-F).

Table 4: Dataset for performance characterization

Videos	**60% training**	**40% testing**	**Total**
Negative videos	259,050 f	172,700 f	431,750 f
Copied segments	16,200 f	10,800 f	27,000 f
			458,750 f

For matching, we have compared the feature vectors with the cosine similarity $SC(X, Y)$ (with two vectors X and Y). It is a common measure for matching of CNN features that is time-efficient and robust [3]. With a unit l_2-norm, it can obtained with a single dot product $X \cdot Y$. Considering m and n the size of the training and testing set, the brute-force comparison has a complexity $O(mn)$ (requiring 50.5 B of matching per feature database with total 455 B). This can be achieved in some hours with a time-efficient implementation[5].

We have applied the characterization protocol of [7,13,15] to evaluate the individual performance of 2D CNN features. All the extracted frames from the copied segments have been labelled with the references in the groundtruth. The negative frames have no label. The performance evaluation has been computed with the P, R and F_1 scores. That is, the maximum cosine similarity will matter and at least one detected frame is required to detect the video.

[4] detailed in the Appendix ...

[5] experiments on a GPU RTX 2070 (7 GiB for the features / 1 GiB for the programs), dataset fully loaded, matching with a fast vector multiplication on all the cores.

3.2 Comparison of 2D CNN features

Based on the dataset and our protocol, we compare here the accuracy of 2D CNN features. Fig. 1 (a) gives the F_1 scores, over a threshold on the cosine similarity, of the different 2D CNN with a common feature extraction method (Last FC). For clarification, the top F_1 scores are reported too in Table 5.

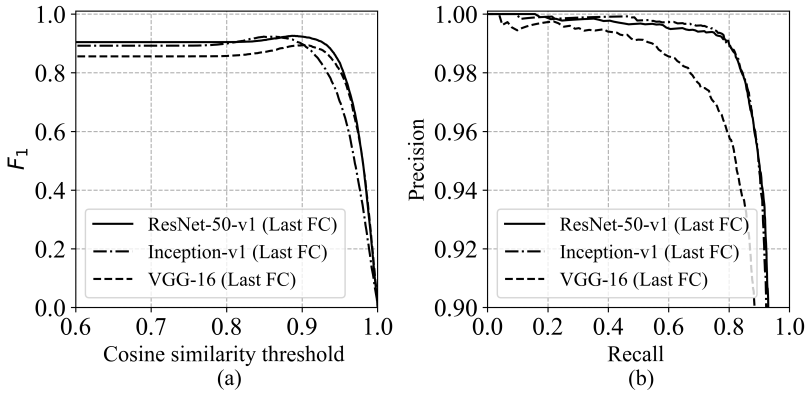

Fig. 1: Comparison of 2D CNN with the Last FC (a) F_1 (b) P/R

The separability for the detection is not achieved even if strong scores are obtained. A maximum of $F_1 \simeq 0.93$ is performed with the ResNet50-v1 network. The different networks present competitive results with a maximum gap of $F_1 \simeq 0.03$. These results are consistent with previous comparisons of 2D CNN in the state-of-the-art [5]. For further analysis, Fig. 1 (b) provides the P/R plot. All the 2D CNN maintain a strong precision at a high level of recall.

Table 5: Comparison of feature extraction methods with the 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

For a comparison of the feature extraction methods, Table 5 gives the top F_1 scores of the different 2D CNN with the Last FC, MAC and R-MAC. For VGG-16, MAC and R-MAC outperform the Last FC method with a slight gap of $F_1 \simeq 0.03$. These methods provide a performance degradation for ResNet50-v1 and Inception-v1 up to a gap of $F_1 \simeq 0.18$. This can be mainly explained by the larger sizes of convolution layers in the VGG-16 network compared to ResNet50-v1 and Inception-v1. This leads more accurate localizations with the MAC and R-MAC features. An equivalent conclusion is also reported in [2].

3.3 Characterization of key-frames with 2D CNN features

The selection of 2D CNN features has a performance impact. However, another important aspect is the ability of video content to be characterized by these features. Indeed, the characterization protocol for PVCD [7,13,15] looks for the maximum cosine similarity between video frames where at least one "good" key-frame is required to detect a video. However, key-frames Fig. 2 with a high-level of noise (a), near-constant (b) or almost duplicate (c) could be difficult to detect. A quantitative analysis of the goodness of key-frames must be established and the regular metrics (P, R and F_1) are misleading on the task. We will investigate this aspect here by providing a characterization protocol of key-frames with 2D CNN features. The goal is to evaluate the performance accuracy of 2D CNN features when facing a large variability of key-frames for PVCD.

(a) (b) (c) (d) (e)

Fig. 2: Examples of key-frames
(a) blurred (b) near-constant (c) almost-duplicate
(d) foreground / background (e) symmetrical

For the needs of characterization, we propose the goodness criterion of Eq. (1). This criterion maximizes the intra and interclass similarity. X is the 2D CNN feature of a positive frame and $\{\tilde{X}_1, \ldots, \tilde{X}_m\}$ its corresponding near duplicate. $\{Y_1, \ldots, Y_{n_1}\}$ is the set of negative 2D CNN features and $\{X_1^*, \ldots, X_{n_2}^*\}$ the positive ones obtained from the other references. $SC_{\min}$ and $SC_{\max}$ are operators to get the minimum and maximum SC between the template X and feature sets. That is, $\phi(X)$ is defined[6] $\in [-1, 1]$ and $\phi(X) > 0$ guaranties a separability[7].

$$\phi(X) = SC_{\min}(X, \{\tilde{X}_1, \ldots, \tilde{X}_m\}) - SC_{\max}(X, \{Y_1, \ldots, Y_{n_1}\}, \{X_1^*, \ldots, X_{n_2}^*\}) \quad (1)$$

Every frame X and its near-duplicates $\{\tilde{X}_1, \ldots, \tilde{X}_m\}$ are aligned with a timestamp t having a precision of $\frac{1}{30}$ second Table 3. The overall set of frames can be modelled with time series Fig. 3. In these series, the $z_1, \ldots, z_{m+1}$ values are derived from $\phi(X)$. For a given frame X at t, we have $z_1 = \phi(X)$, $z_2 = \phi(\tilde{X}_1)$, $\ldots$, $z_{m+1} = \phi(\tilde{X}_m)$. These values can be characterized with statistics (the minimum $z_{\min}$, mean $\overline{z}$ and maximum $z_{\max}$ values of $z_1, \ldots, z_{m+1}$ and their standard deviation σ) and a rate τ accounting the amount of positive criteria.

[6] The Eq. (1) is defined for $SC(X, Y) \in [0, 1]$ with 2D CNN using a RELU function.
[7] No possibility for X to be classified as a false negative (X matched with a negative frame or assigned to another video reference).

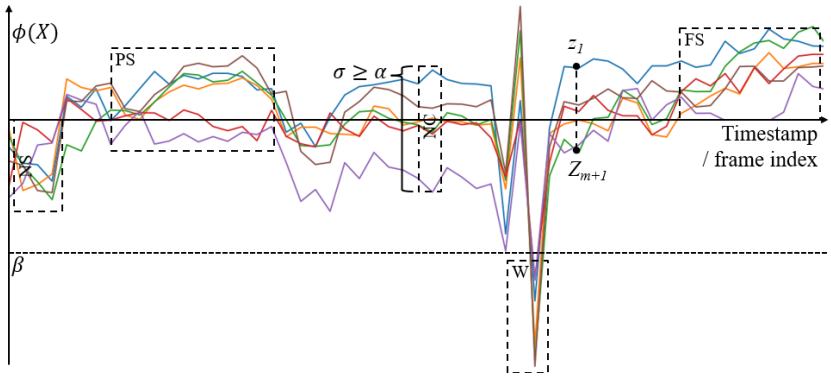

Fig. 3: Modelling with time series

From statistics ($z_{\min}$, $\overline{z}$, $z_{\max}$, σ) and rates τ, the frames can be categorized as detailed in Table 6 and illustrated in Fig. 3. The statistics and rates are compared to thresholds α, β obtained with automatic methods as detailed thereafter. The large variably between the 2D CNN features of a given frame can be detected when an outlier σ value appears greater than the threshold α. This constitutes the set of not consistent frames labelled NC. The frames where the separability cannot be obtained with the 2D CNN features are categorized when $z_{\max} < 0$ then $\tau = 0$. They are labelled NS. From the NS frames, some worst frames labelled W can be filtered out such as $z_{\max} < \beta$. The frames where a partial or fully separability could be obtained with the 2D CNN features are categorized when $\tau \in]0, 1[$ and $\tau = 1$, respectively. They are labelled FS and PS.

Table 6: Categorization of frames

Category	σ	$z_{\min}$	$z_{\max}$	τ
Not Consistent (NC)	$> \alpha$	$\in [-1, 1]$		$\in [0, 1]$
Worst (W)	$\leq \alpha$	$\in [-1, \beta[$		$= 0$
Not Separable (NS)		$\in [\beta, 0[$		
Partially Separable (PS)		< 0	≥ 0	$\in]0, 1[$
Fully Separable (FS)		≥ 0		$= 1$

Table 7 reports the results of categorization on the training set Table 3. We have applied as thresholds $\alpha = 0.05$ and $\beta \in [-0.4, 0]$ obtained with automatic methods detailed thereafter. For the experiments, we have extended the number of positive frames from $16,200$ to $486,000$ with a sampling at the full FPS $= 30$. We have used the VGG-16 with the MAC feature extraction method for tradeoff between a strong detection score $F_1 \simeq 0.92$ Table 5 and the memory constraint. With m and n the numbers of positive and negative frames, the Eq. (1) has a complexity[8] $O(m\left(\frac{m+1}{2}\right) + mn)$. This requires $\simeq 244$ B of matching.

[8] With $S(X, X^*) = S(X^*, X)$, the comparison number of m features is $m\left(\frac{m+1}{2}\right)$.

Table 7: Categorization results of the training set at full FPS= 30

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 %
	21.9 %			**78.1 %**	

A total of $50,844$ timestamps / indices have been obtained Table 3. $\simeq 22\%$ of frames have been categorized as not consistent NC and worst W. Within the remaining $\simeq 78\%$, only $\simeq 13\%$ fit with the partial PS or full separability FS. That is, only a very small amount of "good" key-frames appears in the several videos corresponding to the categories PS and FS. $\simeq 87\%$ of key-frames are hard to detect from their 2D CNN features not consistent or little discriminant.

The categorization results of applied thresholds $\alpha = 0.05$ and $\beta \in [-0.4, 0]$. They must be selected carefully, we have fixed them with automatic methods illustrated in Fig. 4. Fig. 4 (a) plots the cumulative distribution of σ over the $50,844$ indices. The threshold $\alpha \simeq 0.05$ can be easily obtained with an automatic elbow detection. For clarification, the cumulative rate of indices with $\tau = 0$ (over all the indices $\tau \in [0, 1]$) is given for $\sigma > \alpha$. $\ll 1\%$ of indices have a $\tau \neq 0$. The threshold β has been fixed to detect outliers for indices with $\tau = 0$ and $\sigma \leq \alpha$ reference per reference. Fig. 4 (b) illustrates the method. For each reference, a mean $\overline{\mathcal{Z}}$ of indices is computed. This mean serves to fix the threshold $\beta = \overline{\mathcal{Z}}$. The indices with $z_{\max} < \overline{\mathcal{Z}}$ are categorized as worst frames W. Considering the 243 references Table 3, we have obtained a range $\beta \in [-0.4, 0]$.

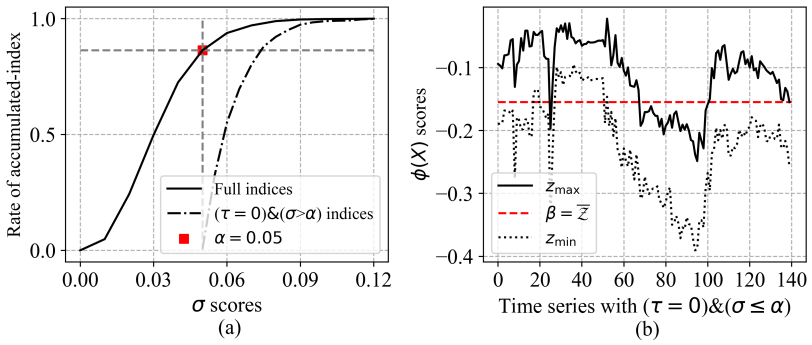

Fig. 4: (a) distribution of σ (for α) (b) times series with $\tau = 0$ and $\sigma \leq \alpha$ (for β)

Fig. 2 provides examples of key-frames for the different categories. Fig. 2 (d, e) gives key-frames labelled FS containing distinguished shapes (e.g. background / foreground text). They are easy to detect with 2D CNN features [20]. However, they are difficult to catch from videos as they constitute only $\simeq 3\%$ of the total amount of key-frames Table 7. Fig. 2 (b, c) gives key-frames having a worst label

W with a near-constant or an altered visual content (e.g. inclusion of logos). Even if they constitute a small part of key-frame $\simeq 8\%$ Table 7, they must be carefully avoided for PVCD. Fig. 2 (a) shows a key-frame with a high level of blurring labelled NC. Such key-frames have 2D CNN features with a large variability and little discriminant. They are hard to detect [14]. At last, $\simeq 65\%$ of key-frames are categorized as NS. The 2D CNN features of these key-frames cannot be detected efficiently.

4 Conclusions and perspectives

Based on a large-scale video dataset, this paper gives a performance characterization of 9 common 2D CNN features used for PVCD. The experiments have been driven on 4.4 M feature vectors with 700 B of comparisons. The separability is not achieved on the detection problem even if strong scores are obtained with a maximum of $F_1 \simeq 0.93$. The different networks present competitive results with a maximum gap of $F_1 \simeq 0.03$. As a general trend, features extracted from recent 2D CNN such as ResNet50 perform better. A correlation appears between the feature extraction methods and the 2D CNN architectures (e.g. VGG-16 with the MAC and R-MAC features). These different conclusions are consistent with the state-of-the-art in the computer vision field.

From 2D CNN features modelled as time series, a method for categorization of key-frames is proposed. This method allows a deeper characterization of a PVCD problem with 2D CNN features. It provides (**i**) a fine categorization of key-frames (**ii**) a characterization of 2D CNN features for separability and consistency (**iii**) a quantitative analysis of the goodness of key-frames. It highlights the performance limits of 2D CNN features when facing blurred, near-constant or almost-equivalent key-frames. In addition, a large part of key-frames ($\simeq 87\%$) cannot be classified efficiently from 2D CNN features. These limitations will be explored in our future works by investigating the robust key-frame selection and learning of 2D CNN features to further improve the PVCD performance.

References

1. Cheng, H., Wang, P., Qi, C.: Cnn features based unsupervised metric learning for near-duplicate video retrieval. In: Open-access repository (arXiv). No. 2105.14566v1 (2021)
2. Cools, A., Belarbi, M., Mahmoudi, S.: A comparative study of reduction methods applied on a convolutional neural network. Electronics **11**, 1422 (2022)
3. Gkelios, S., Sophokleous, A., Plakias, S., Boutalis, Y., Chatzichristofis, S.: Deep convolutional features for image retrieval. Expert Systems With Applications **177**(114940) (2021)
4. Han, Z., He, X., Tang, M., LV, Y.: Video similarity and alignment learning on partial video copy detection. In: ACM International Conference on Multimedia (MM). pp. 4165–4173 (2021)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on computer vision and pattern recognition (CVPR). pp. 770–778 (2016)

6. He, S., He, Y., Lu, M., Jiang, C., Yang, X., Qian, F., Zhang, X., Yang, L., Zhang, J.: Transvcl: Attention-enhanced video copy localization network with flexible supervision. In: AAAI Conference on Artificial Intelligence (AAAI) (2023)
7. He, S., Yang, X., Jiang, C., Liang, G., Zhang, W., Pan, T., Wang, Q., Xu, F., Li, C., Liu, J., et al.: 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 (2022)
8. Jiang, C., Huang, K., He, S., Yang, X., Zhang, W., Zhang, X., Cheng, Y., Yang, L., Wang, Q., Xu, F.: Learning segment similarity and alignment in large-scale content based video retrieval. In: ACM International Conference on Multimedia (MM). pp. 1618–1626 (2021)
9. Jiang, Q., He, Y., Li, G., Lin, J., Li, L., Li, W.: Svd: A large-scale short video dataset for near-duplicate video retrieval. In: International Conference on Computer Vision (ICCV). pp. 5281–5289 (2019)
10. Jiang, Y., Wang, J.: Partial copy detection in videos: A benchmark and an evaluation of popular methods. IEEE Transactions on Big Data **2**(1), 32–42 (2016)
11. Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, I.: Fivr: Fine-grained incident video retrieval. IEEE Transactions on Multimedia **21**(10), 2638–2652 (2019)
12. Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, Y.: Near-duplicate video retrieval with deep metric learning. In: International Conference on Computer Vision Workshops (ICCV). pp. 347–356 (2017)
13. Le, V., Delalandre, M., Conte, D.: A large-scale tv dataset for partial video copy detection. In: International Conference on Image Analysis and Processing (ICIAP). Lecture Notes in Computer Science (LNCS), vol. 13233, pp. 388–399 (2022)
14. Roy, P., Ghosh, S., Bhattacharya, S., Pal, U.: Effects of degradations on deep neural network architectures. In: Open-access repository (arXiv). No. 1807.10108v5 (2023)
15. Tan, W., Guo, H., Liu, R.: A fast partial video copy detection using knn and global feature database. In: Winter Conference on Applications of Computer Vision (WACV). pp. 2191–2199 (2022)
16. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. In: International Conference on Learning Representations (ICLR). pp. 1–12 (2016)
17. Wang, K., Cheng, C., Chen, Y., Song, Y., Lai, S.: Attention-based deep metric learning for near-duplicate video retrieval. In: International Conference on Pattern Recognition (ICPR). pp. 5360–5367 (2021)
18. Wang, L., Bao, Y., Li, H., Fan, X., Luo, Z.: Compact cnn based video representation for efficient video copy detection. In: International conference on multimedia modeling (MMM). pp. 576–587 (2017)
19. Zhang, C., Hu, B., Suo, Y., Zou, Z., Ji, Y.: Large-scale video retrieval via deep local convolutional features. Advances in Multimedia **2020**, 1687–5680 (2020)
20. Zhang, X., Gao, J.: Measuring feature importance of convolutional neural networks. IEEE Access **8**, 196062–196074 (2020)
21. Zhang, X., Xie, Y., Luan, X., He, J., Zhang, L., Wu, L.: Video copy detection based on deep cnn features and graph-based sequence matching. Wireless Personal Communications **103**(1), 401–416 (2018)
22. Zhao, G., Zhang, B., Zhang, M., Li, Y., Liu, J., Wen, J.: Star-gnn: spatial-temporal video representation for content-based retrieval. In: International Conference on Multimedia and Expo (ICME). pp. 01–06 (2022)

Appendix

We present in this appendix the pipeline to extract frames from the videos and groundtruth of the STVD dataset. The STVD dataset is constituted of six test sets A to F having different sources of degradation (e.g. pixel attack, video speeding). We have selected the test set D, illustrated in the next figure, related to scalability and global transformations. As detailed in the next Table, it includes $3,213$ and $12,165$ positive and negative videos having a total and a mean duration of $1,960$ hours and 7.5 minutes, respectively. The videos are encoded at 30 FPS with a controlled quality[9]. The source of video capture ensures a low contrast variation[10]. Global transformations have been applied and combined including flipping, rotation and inclusion of black borders. This test set fits well with the 2D CNN features that are translation, scale and rotation invariant.

Global transformations in the test set D of the STVD dataset

Data pipeline for frame extraction

Videos	**Number**	**Duration (h)**	**FPS**	**60% training**	**40% testing**	**Total**
Negative videos	12,165	1,545 h	$\simeq 0.08$	259,050 f	172,700 f	431,750 f
Positive videos	3,869	415 h				
Copied segments	4,436	7.5 h	1	16,200 f	10,800 f	27,000 f

458,750 f $\simeq$ 7 GiB of 4,096-F features

We have randomly split the videos into a training and a testing set as detailed in the Table. The rates of 60% and 40% have been applied. We have used the $12,165$ negative videos without any modification. We have re-generated and added 656 videos to the $3,213$ positive videos. Indeed, the 243 references have occurrences from 1 up to 167. To fit with the splitting process, a minimum of 10 occurrences per reference is needed. We have then used the timestamps information to extract the copied segments. As detailed in [13], these segments have a duration $\in [1; 25]$ seconds and one and more segments could appear in a video. We have obtained $4,436$ copied segments having a duration of 7.5 hours.

[9] 144×192 pixels per frame at 28 kbps characterized with MSE in [13]
[10] characterized with a Contrast Noise Ratio (CNR) in [13]

We have sampled the negative videos and copied segments with static FPS to get frames. Static FPS is a common method for key-frame selection. As detailed in the Table, a FPS= 1 has been applied to the copied segments. We have obtained $27,000$ feature vectors split into the training and testing sets. For the negative videos, we have fixed a FPS $\simeq 0.08$ for tradeoff between the scalability and the memory constraint. We have considered for the experiments a GPU with a 8 GiB of memory (e.g. *Nvidia* GPU RTX 2070). In such a GPU, 7 GiB can be allocated to the features[11] for a total amount of $458,750$ considering a maximum size of $4,096$-F. We have extracted $431,750$ features vectors from the negative videos dispatched into the training and testing sets.

[11] 1 GiB to store the program, frameworks and for the results.