

# TRÍCH CHỌN ĐẶC TRƯNG TỪ CÁC MẠNG NƠ-RON TÍCH CHẬP CHO BÀI TOÁN PHÁT HIỆN SAO CHÉP MỘT PHẦN VIDEO

Lê Văn Hào<sup>1</sup>, Lê Đình Nghiệp<sup>2</sup>, Nguyễn Văn Cường<sup>1</sup>

## TÓM TẮT

*Các mạng 2D nơ-ron tích chập là những thành phần chính trong hệ thống phát hiện sao chép một phần video. Chúng đóng vai trò quan trọng trong nhiệm vụ truy hồi và đối sánh video trên một cơ sở dữ liệu lớn. Tuy nhiên, đặc tính hiệu năng của các phương pháp trích chọn đặc trưng này ít được nghiên cứu. Bài báo này sẽ đóng góp hai phần chính. Thứ nhất, chúng tôi cung cấp một thí nghiệm trên bộ dữ liệu quy mô lớn để chứng minh tính phổ quát và làm rõ đặc trưng hiệu năng của các mạng nơ-ron phổ biến. Kế đến, chúng tôi đề xuất một phương pháp theo mô hình chuỗi thời gian để chỉ ra ưu và nhược điểm của các đặc trưng ảnh trích xuất từ các mạng nơ-ron trong bài toán phát hiện sao chép một phần video.*

**Từ khóa:** Mạng nơ-ron tích chập, trích chọn đặc trưng, sao chép một phần video.

**DOI:** <https://doi.org/10.70117/hdujs.79.09.2025.828>

## 1. ĐẶT VẤN ĐỀ

Phát hiện sao chép một phần video (Partial Video Copy Detection - PVCD) mục tiêu là tìm các đoạn video nhỏ đã bị biến đổi từ một video tham chiếu gốc trong một cơ sở dữ liệu lớn. Chủ đề PVCD gần đây được nhiều sự quan tâm trong nghiên cứu cũng như trong các ứng dụng thực tế [10][21]. Tuy nhiên, theo hiểu biết của chúng tôi thì vấn đề hiệu năng của từng loại mạng nơ-ron và đánh giá ưu, nhược điểm của từng phương pháp trích chọn đặc trưng là ít được quan tâm trong các tài liệu y văn hiện có [14].

Để giải quyết vấn đề trên, trong bài báo này, chúng tôi đóng góp 2 phần chính. Thứ nhất, dựa trên một cơ sở dữ liệu lớn về PVCD, chúng tôi đã thực hiện các thí nghiệm với hơn 700 tỷ phép đối sánh của hơn 4,4 triệu vec-tơ đặc trưng. Những thí nghiệm này đã chỉ ra và lý giải những đặc trưng mang tính quyết định ảnh hưởng tới hiệu năng của các hệ thống PVCD sử dụng các 2D CNN. Kế đến, một giao thức thông thường để đánh giá các mạng 2D CNN ở mức video thường đem đến sự hiểu nhầm cho rằng hiệu năng của các hệ thống PVCD đã đạt được các kết quả hoàn hảo. Do đó, chúng tôi đề xuất một phương pháp cho phép phân tích sâu và mịn hơn ở mức các khung hình chính (key-frame) thay vì mức video như các nghiên cứu trước đã chỉ ra.

Bài báo này là phiên bản dịch và chỉnh sửa từ công trình nguyên thủy của nhóm tác giả được công bố trên [14]. Chúng tôi thực hiện với sự cho phép của tác giả gốc, nhằm mục đích làm cho nghiên cứu có thể tiếp cận được với nhiều đối tượng hơn. Mọi nỗ lực đã được thực hiện để đảm bảo tính chính xác và trung thực của nội dung gốc trong khi điều chỉnh nó cho độc giả của tạp chí này để phù hợp với khuôn mẫu và định dạng. Bài báo được cấu trúc

<sup>1</sup> Khoa Kỹ thuật, Công nghệ và Truyền thông, Trường Đại học Hồng Đức; Email: levanhao@hdu.edu.vn

<sup>2</sup> Trung tâm Công nghệ thông tin, Truyền thông và Thư viện, Trường Đại học Hồng Đức

các phần như sau. Phần đầu cung cấp những khái niệm cơ bản về PVCD và đóng góp chính của chúng tôi. Kế đến, chúng tôi đưa ra một phân tích triệt để về các nghiên cứu gần đây liên quan đến PVCD sử dụng các mạng 2D CNN. Các kết quả nghiên cứu được trình bày ở phần 3. Cuối cùng là các kết luận và một số hướng nghiên cứu trong tương lai.

## 2. PHƯƠNG PHÁP NGHIÊN CỨU

### 2.1. Nội dung và phạm vi nghiên cứu

Ứng dụng 2D CNN trong chủ đề PVCD là rất đa dạng và phong phú [11][12][15][17]. Tuy nhiên, hầu hết các nghiên cứu trước đây mới chỉ tập trung vào việc đối sánh ở mức toàn cục-video. Vì thế, trong bài báo này nhóm tác giả tập trung nghiên cứu và phân tích phương pháp phát hiện sao chép ở mức khung hình sử dụng 2D CNN để đưa ra những phát hiện mới và chi tiết hơn so với các công trình đã được công bố.

Bảng 1 dưới đây trình bày các nghiên cứu tiêu biểu chính về các hệ thống PVCD sử dụng 2D CNN. Đầu tiên là trích xuất các khung hình chính bằng các kỹ thuật như lấy cố định FPS. Các khung hình gần nhau có độ tương tự cao, để tránh dư thừa và tốn chi phí lưu trữ, một số phương pháp trích chọn khung hình chính cũng được đề xuất [1][20].

**Bảng 1. Tổng quan hệ thống PVCD sử dụng 2D CNN**

|                       |                                                                                                                        |
|-----------------------|------------------------------------------------------------------------------------------------------------------------|
| Chọn khung hình chính | Sử dụng FPS [2,4,6,7,9,10-13,16,18,19,23]; Phương pháp khác [1,20]                                                     |
| Chọn mô hình mạng     | VGG [9,11,13,16,19,20,22,23]; ResNet [4,7,8,11,16]; InceptionNet [1,11,12,18]; AlexNet [1,10,12,22]                    |
| Trích chọn đặc trưng  | Tầng kết nối đầy đủ [1,10-13,19,22]; Tầng tích chập [1,4,7,8,11,12,16,18,20,23]; PCA [4,6,19,20]; Phân tích RoI [8,23] |
| Đối sánh video        | Mức khung hình [7,13,16]; Mức video [1,4,6,7,8,16,22]                                                                  |

Các khung hình chính sau đó được đưa qua mạng đã huấn luyện (pre-trained) 2D CNN như AlexNet, VGG, ResNet... để trích chọn các đặc trưng. Thông thường các mạng này xử lý ảnh màu RGB với một kích thước đầu vào cố định. Chúng cũng có những phiên bản khác nhau và kiến trúc của mỗi mạng cũng đưa ra những hiệu năng khác nhau.

Các đặc trưng của hình ảnh được trích xuất thông qua hai phương pháp chính. Một là trích xuất từ lớp kết nối đầy đủ cuối cùng, ngoại trừ lớp dành cho phân loại ảnh (Fully Connected - FC). Hoặc đặc trưng ảnh có thể trích chọn từ những tầng tích chập. Đối với phương pháp trích xuất từ tầng tích chập, hai kỹ thuật phổ biến được áp dụng đó là lấy giá trị cực đại theo tầng (MAC) hoặc lấy giá trị cực đại theo vùng (Maximum Activations of Convolutions (MAC) và Regional-MAC (R-MAC) [17].

Cuối cùng, các vec-tơ đặc trưng được đem đi đối sánh để xác định liệu rằng có phải nó được sao chép từ video tham chiếu hay không. Phương pháp thứ nhất là phát hiện sao chép dựa trên kết quả đối sánh từng khung hình đơn. Giả thiết đưa ra là 2 video được coi là sao chép nếu ít nhất có một khung hình được phát hiện trong nhãn (groundtruth). Phương pháp thứ hai là đối sánh ở mức toàn cục thông qua một ma trận là kết quả độ tương tự giữa các khung hình theo chiều thời gian. Trong cả 2 phương pháp, vec-tơ đặc trưng ảnh thường được chuẩn hóa theo chuẩn L2 trước khi áp dụng tính độ tương tự Cosine. Nhằm tối ưu thời gian đối sánh và không gian lưu trữ, một số phương pháp giảm chiều kích thước vec-tơ cũng được áp dụng như Pooling hay PCA.

## 2.2. Phương pháp nghiên cứu

Trong bài báo này nhóm tác giả đã áp dụng và triển khai quy trình nghiên cứu gồm hai phương pháp. Đầu tiên, phương pháp nghiên cứu lý thuyết nhằm hệ thống hóa kiến thức về các kỹ thuật và thuật toán liên quan đến trích chọn đặc trưng ảnh từ mạng nơ-ron cho bài toán PVCD. Kế đến, phương pháp nghiên cứu thực nghiệm thông qua việc tổ chức chuẩn bị dữ liệu, cài đặt thuật toán trên các bộ dữ liệu chuẩn, nhằm để so sánh kết quả nghiên cứu liên quan. Dựa trên kết quả thực nghiệm, chúng tôi chỉ ra và luận giải được những hạn chế của các phương pháp hiện tại, cũng như làm nổi bật phương pháp đề xuất của chúng tôi.

## 3. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

### 3.1. Bộ dữ liệu và các bước xử lý

Để đánh giá hiệu năng các hệ thống PVCD, một bộ cơ sở dữ liệu cần được lựa chọn để thực hiện các thí nghiệm. Trong y văn gần đây, một vài bộ dữ liệu được công bố dành cho phát hiện sao chép video và truy vấn các video gần trùng lặp. Bảng 2 tóm tắt một số đặc trưng chính của các bộ dữ liệu đã được đề xuất.

**Bảng 2. Các bộ dữ liệu được sử dụng trong đánh giá hiệu suất hệ thống PVCD**

| Bộ dữ liệu       | VCDB    | SVD      | STVD     | VCSL     |
|------------------|---------|----------|----------|----------|
| Bài báo          | [10]    | [9]      | [13]     | [7]      |
| Cách biến đổi    | Thật    | Tổng hợp | Tổng hợp | Thật     |
| Thời lượng (h)   | 2,030 h | 197 h    | 10,660 h | 17,416 h |
| Video tham chiếu | 28      | 1,206    | 243      | 122      |
| Cấp dương tính   | 9 K     | -        | 1,688 K  | 281 K    |
| Độ chính xác     | 1 s     | -        | 1/30 s   | 1 s      |

Để thực hiện các thí nghiệm, STVD đã được lựa chọn do một số đặc tính quan trọng như (1) bộ dữ liệu được thu thập trên các tín hiệu từ TV với độ nhiễu rất nhỏ để cho phép điều chỉnh dễ dàng các phương pháp biến đổi hình ảnh và video, (2) STVD là bộ dữ liệu có kích thước lớn nhất trong y văn hiện nay, (3) cơ sở dữ liệu này cũng cung cấp một phân phối cân bằng giữa video dương tính và âm tính, (4) nhãn dữ liệu của STVD chính xác tới mức khung hình so với các bộ cơ sở dữ liệu khác.

Từ bộ cơ sở dữ liệu STVD chúng tôi thực hiện một số bước tiền xử lý chuẩn để trích xuất tổng cộng 458,750 khung hình được mô tả trong Bảng 3. Bộ dữ liệu được chia làm hai tập con gồm tập cơ sở gốc và tập kiểm thử. Chúng tôi lựa chọn 3 kiến trúc mạng nơ-ron phổ biến nhất gồm VGG-16, ResNet-50 và Inception-v1. Về phương pháp trích xuất đặc trưng ảnh chúng tôi sử dụng 3 phương pháp chính gồm Last FC, MAC và R-MAC như đã giới thiệu ở trên. Các đặc trưng ảnh sau khi trích xuất từ mạng nơ-ron được chuẩn hóa về dạng L2 với các kích thước 512, 1024, 2048 và 4096 chiều.

**Bảng 3. Phân chia tập dữ liệu STVD**

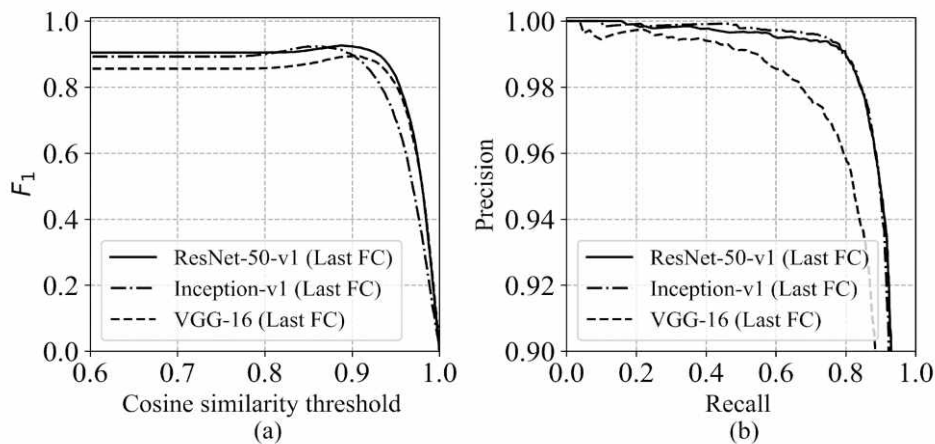
| Video          | 60% tập cơ sở gốc | 40% kiểm thử | Tổng           |
|----------------|-------------------|--------------|----------------|
| Video giả      | 259,050           | 172,700      | 431,750        |
| Video sao chép | 16,200            | 10,800       | 27,000         |
|                |                   |              | <b>458,750</b> |

Bước đối sánh ảnh được song song hóa hiệu quả trên bộ xử lý đồ họa (GPU) sử dụng độ đo tương tự Cosine giữa hai vec-tơ đặc trưng ảnh [3]. Trong đó, khi hai vec-tơ được chuẩn hóa về dạng L2, độ tương tự Cosine giữa hai vec-tơ tương đương với phép tích vô hướng giữa hai vec-tơ này.

Để kiểm chứng đặc tính hiệu năng, chúng tôi sử dụng các độ đo chuẩn gồm độ chính xác P (Precision), độ phủ R (Recall), và độ trung hòa  $F_1$ . Trong đó, độ tương tự Cosine lớn nhất được sử dụng để xác định 2 hình ảnh có là bản sao chép của nhau hay không.

### 3.2. So sánh giữa các mô hình mạng 2D CNN

Dựa trên bộ dữ liệu và qui trình xử lý dữ liệu, hiệu năng của các kiến trúc mạng 2D CNN được mô tả trong Hình 1 và trình bày chi tiết trong Bảng 4.



Hình 1. So sánh 2D CNN với Last FC (a)  $F_1$  (b) P/R

Qua thực nghiệm, chúng tôi thấy rằng độ ảnh hưởng của các kiến trúc mạng khác nhau là không tách biệt đáng kể (một khoảng cách nhỏ 0.03) mặc dù điểm  $F_1$  cao nhất thu được với kiến trúc mạng ResNet-50 ở mức gần 0.93. Trong Hình 1 (b) cho thấy các mạng 2D CNN vẫn duy trì độ chính xác ở mức cao cùng với độ phủ lớn.

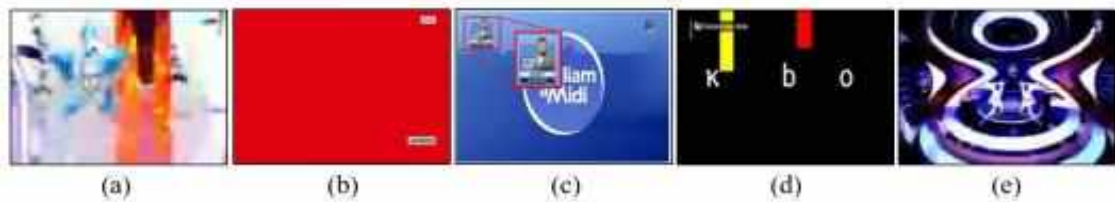
Bảng 4. Danh sách các điểm  $F_1$  của 9 loại đặc trưng 2D CNN

|              | Last FC      | MAC          | R-MAC |
|--------------|--------------|--------------|-------|
| ResNet50-v1  | <b>0.926</b> | 0.828        | 0.823 |
| Inception-v1 | <b>0.923</b> | 0.738        | 0.782 |
| VGG-16       | 0.894        | <b>0.922</b> | 0.918 |

Về phương pháp trích xuất đặc trưng, chúng tôi nhận thấy một kiến trúc mạng VGG dùng MAC và R-MAC cho một hiệu suất tốt hơn  $F_1 = 0.03$  so với phương pháp Last FC. Tuy nhiên, đối với kiến trúc mạng ResNet và InceptionNet thì chúng tôi thu được một khoảng cách lớn  $F_1=0.18$  giữa trích xuất từ tầng tích chập và trích xuất từ tầng kết nối đầy đủ. Điều này được giải thích do kích thước tầng cuối cùng VGG lớn hơn gấp hai lần (14x14) so với các kiến trúc ResNet và InceptionNet (7x7 hoặc 8x8). Điều này dẫn đến việc áp dụng phép toán cực đại lên các vùng khác nhau của bề mặt lớp tích chập trong VGG đem lại một độ chính xác cao hơn.

### 3.3. Độ ảnh hưởng tới hiệu suất của các đặc trưng 2D CNN

Việc lựa chọn các kiến trúc mạng và phương pháp trích chọn đặc trưng có một số ảnh hưởng tới hiệu suất hệ thống PVCD đã được chúng tôi chỉ ra ở phần 3.2. Tuy nhiên, việc xử lý video trong chủ đề PVCD đem lại một khía cạnh khác liên quan trực tiếp tới nội dung của video thay vì xử lý các hình ảnh đơn. Thật vậy, trong các nghiên cứu [7, 13, 16] tập trung vào độ đo tương tự Cosine cực đại giữa hai khung hình làm điều kiện quyết định xem video đó có bị sao chép hay không. Trong phần này chúng tôi sẽ chỉ ra những điểm hạn chế của đặc trưng sử dụng 2D CNN khi đối mặt với các độ nhiễu cao, nội dung ảnh chỉ gồm hầu như một màu đồng nhất, hay các ảnh gần trùng lặp về nội dung. Chúng được miêu tả trong Hình 2 (a) (b) và (c) dưới đây.



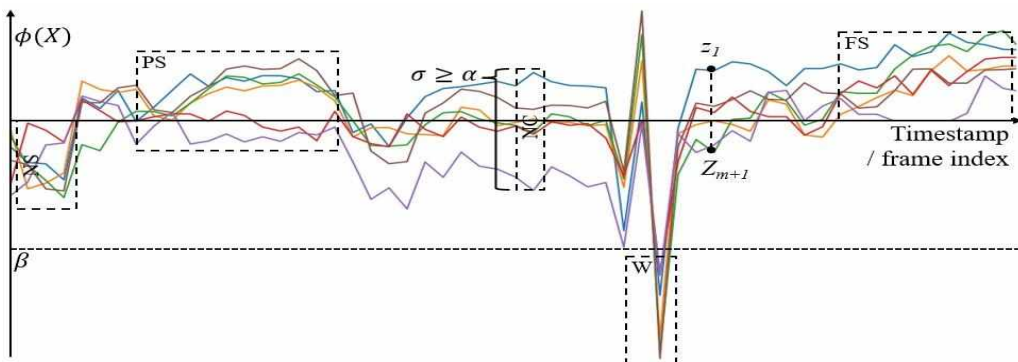
**Hình 2. Một số ví dụ về khung hình chính**

*Ghi chú: Khung hình bị mờ (b) khung hình với một màu đồng nhất (c) khung hình có nội dung gần trùng lặp (d) khung hình có nội dung tách biệt nền/màu (e) khung hình có nội dung được bố cục dưới dạng lưới*

Chúng tôi đề xuất trong bài báo này một độ đo ‘tiêu chuẩn’ để xác định tính phân biệt giữa các đặc trưng trong khung hình. Độ đo  $\phi(X)$  được xác định là điểm lớn nhất giữa độ tương tự bên trong lớp chứa video dương tính và giữa các lớp khác với nhau. Một kết quả lớn hơn không (0) là được cho đảm bảo khả năng phân loại tốt. Công thức dưới đây mô tả độ đo chuẩn được xác định.

$$\phi(X) = SC\_min(A, \{\tilde{A}_1, \dots, \tilde{A}_m\} - SC\_max(A, \{Y_1, \dots, Y_{n1}\}, \{\tilde{a}_1, \dots, \tilde{a}_{n2}\})$$

Mỗi một khung hình  $X$  và bản sao của nó  $\{\tilde{A}_1, \dots, \tilde{A}_m\}$  là được căn chỉnh theo chuỗi thời gian  $t$  với độ chính xác 1/30 giây. Chúng tôi sẽ dựa trên mô hình chuỗi thời gian này và độ đo ‘tiêu chuẩn’ để thực hiện các phân tích với các giá trị thống kê như min, mean, max, standard deviation. Mô hình hóa theo chuỗi thời gian được mô tả như Hình 3 dưới đây.



**Hình 3. Mô hình hóa theo chuỗi thời gian**

*Ghi chú: Trục tung là biểu thị các giá trị của  $\phi(X)$ , trong khi trục hoành là biểu thị dấu giá trị theo khung thời gian  $t$ . Các kí hiệu NC, NS, W, PS, FS là được giải thích ở Bảng 5*

Từ mô hình này, chúng tôi chia các khung hình chính thành các loại khác nhau được mô tả trong Hình 2 và viết tắt trong Hình 3. Ví dụ, Hình 2 (d, e) minh họa các khung hình được gán nhãn FS chứa các nội dung có mức độ phân biệt cao (bố cục lưới, phân biệt màu nền / nội dung). Những khung hình FS là dễ dàng được phát hiện với 2D CNN. Tuy nhiên, chúng chỉ chiếm 3% tổng số lượng khung hình có trong video. Bảng 5 thống kê tổng số lượng khung hình và phân loại của chúng. Hình 2 (b, c) minh họa các khung hình với nhãn W miêu tả các hình với màu đồng nhất hoặc gần trùng lặp (chèn logo, chữ). Mặc dù chỉ chiếm một phần nhỏ với gần 8%, nhưng chúng phải được loại bỏ một cách cẩn thận. Hình 2 (a) miêu tả những khung hình bị mờ bởi các hiệu ứng chuyển khung cảnh trong video được nhãn NC. Những khung hình như thế có một độ biến thiên (độ lệch chuẩn) lớn với mức độ phân biệt khá nhỏ, dẫn đến 2D CNN rất khó để phát hiện sao chép. Cuối cùng, gần 65% số lượng khung hình được gán nhãn NS. Những đặc trưng khung hình chính này khó được phát hiện một cách hiệu quả với các 2D CNN.

**Bảng 5. Bảng thống kê phân loại khung hình chính**

| Số khung hình | 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%  |

*Ghi chú: Trong đó các kí hiệu  $\phi(X)$  là giá trị được tính theo công thức đã trình bày ở trên, NC là dạng viết tắt cho loại khung hình ‘Not Consistent’, W là dạng viết tắt cho ‘Worst’, NS là dạng viết tắt cho ‘Not Separable’, PS là dạng viết tắt cho ‘Partial Separable’, và FS là dạng viết tắt cho ‘Full Separable’.*

### 3.4. Kết luận và hướng nghiên cứu

Dựa trên một cơ sở dữ liệu video lớn, chúng tôi đã thực hiện các thí nghiệm trên 9 loại đặc trưng 2D CNN được dùng trong PVCD. Các thí nghiệm này được đánh giá trên một không gian chứa gần 4.4 triệu vec-tơ đặc trưng với gần 700 tỉ phép đối sánh. Chúng tôi đã chỉ ra một sự tách biệt nhỏ về hiệu suất giữa các kiến trúc mạng khác nhau trong đó điểm số cao nhất là mạng Resnet-50 tại  $F_1=0.93$ . Tiếp theo, chúng tôi cũng thấy rằng các phương pháp trích chọn đặc trưng là có liên quan với các kiến trúc mạng. Cụ thể, đối với mạng VGG thì phương pháp MAC, R-MAC cho hiệu suất cao hơn so với phương pháp Lats-FC với độ khác biệt  $F_1=0.03$ .

Từ mô hình theo chuỗi thời gian, chúng tôi đề xuất một phương pháp để phân loại các khung hình chính dựa trên những đặc trưng của chúng. Phương pháp này cho phép các phân tích sâu và mịn hơn phương pháp toàn cục – xử lý mức video. Qua thí nghiệm, chúng tôi chỉ ra và nhấn mạnh các giới hạn về tính mạnh mẽ của các đặc trưng sử dụng mạng nơ-ron trong các trường hợp video bị mờ (blurring), khung hình với một màu duy nhất, hoặc các khung hình gần trùng lặp. Khi xử lý với video, một lượng lớn các khung hình này (chiếm gần 87%) là rất khó để phát hiện sao chép một cách hiệu quả với 2D CNN. Từ những giới hạn này của 2D CNN chúng tôi hướng đến điều tra một số phương pháp lựa chọn khung hình chính có tính mạnh mẽ và phân hóa cao bằng việc sử dụng các phương pháp huấn luyện lại nhằm cải thiện hơn nữa hiệu suất của PVCD.

## TÀI LIỆU THAM KHẢO

- [1] Cheng, H., Wang, P., Qi, C. (2021), *Cnn features based unsupervised metric learning for near-duplicate video retrieval*. In: Open-access repository (arXiv). No. 2105.14566v1.
- [2] Cools, A., Belarbi, M., Mahmoudi, S. (2022), *A comparative study of reduction methods applied on a convolutional neural network*. Electronics 11, 1422.
- [3] Gkelios, S., Sophokleous, A., Plakias, S., Boutalis, Y., Chatzichristofis, S. (2021), *Deep convolutional features for image retrieval*. Expert Systems With Applications 177(114940).
- [4] Han, Z., He, X., Tang, M., LV, Y. (2021), *Video similarity and alignment learning on partial video copy detection*, In: MM., 4165-4173.
- [5] He, K., Zhang, X., Ren, S., Sun, J. (2016), *Deep residual learning for image recognition*. In: CVPR., 770-778.
- [6] He, S., He, Y., Lu, M., Jiang, C., Yang, X., Qian, F., Zhang, X., Yang, L., Zhang, J. (2023), *Transvcl: Attention-enhanced video copy localization network with flexible supervision*. In: AAAI.
- [7] He, S., Yang, X., Jiang, C., Liang, G., Zhang, W., Pan, T., Wang, Q., Xu, F., Li, C., Liu, J., et al. (2022), *A large-scale comprehensive dataset and copy-overlap aware evaluation protocol for segment-level video copy detection*. In: CVPR., 21086-21095.
- [8] Jiang, C., Huang, K., He, S., Yang, X., Zhang, W., Zhang, X., Cheng, Y., Yang, L., Wang, Q., Xu, F. (2021), *Learning segment similarity and alignment in large-scale content-based video retrieval*. In: MM., 1618-1626.
- [9] Jiang, Q., He, Y., Li, G., Lin, J., Li, L., Li, W. (2019), *Svd: A large-scale short video dataset for near-duplicate video retrieval*. In: ICCV., 5281-5289.
- [10] Jiang, Y., Wang, J. (2016), *Partial copy detection in videos: A benchmark and an evaluation of popular methods*. IEEE Transactions on Big Data 2(1), 32-42.
- [11] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, I. (2019), *Fivr: Finegrained incident video retrieval*, IEEE Transactions on Multimedia 21(10), 2638-2652.
- [12] Kordopatis-Zilos, G., Papadopoulos, S., Patras, I., Kompatsiaris, Y. (2017), *Nearduplicate video retrieval with deep metric learning*. In: ICCV., 347-356.
- [13] Le, V.H., Delalandre, M., Conte, D. (2022), *A large-scale tv dataset for partial video copy detection*. In: ICIAP. vol.13233, 388-399.
- [14] Le, V.H., Delalandre, M., Cardot, H. (2023), *Performance Characterization of 2D CNN Features for Partial Video Copy Detection*, Conference on Computer Analysis of Images and Patterns (CAIP), 205-215.
- [15] Roy, P., Ghosh, S., Bhattacharya, S., Pal, U. (2023), *Effects of degradations on deep neural network architectures*, In: Open-access repository (arXiv). No. 1807.10108v5.
- [16] Tan, W., Guo, H., Liu, R. (2022), *A fast partial video copy detection using knn and global feature database*, In: WACV., 2191-2199.
- [17] Tolias, G., Sicre, R., Jégou, H. (2016), *Particular object retrieval with integral maxpooling of cnn activations*. In: ICLR., 1-12.

- [18] Wang, K., Cheng, C., Chen, Y., Song, Y., Lai, S. (2021), *Attention-based deep metric learning for near-duplicate video retrieval*. In: ICPR., 5360-5367.
- [19] Wang, L., Bao, Y., Li, H., Fan, X., Luo, Z. (2017), *Compact cnn based video representation for efficient video copy detection*. In: MMM., 576-587.
- [20] Zhang, C., Hu, B., Suo, Y., Zou, Z., Ji, Y. (2020), *Large-scale video retrieval via deep local convolutional features*. Advances in Multimedia 2020, 1687-5680.
- [21] Zhang, X., Gao, J. (2020), *Measuring feature importance of convolutional neural networks*. IEEE Access 8, 196062-196074.
- [22] Zhang, X., Xie, Y., Luan, X., He, J., Zhang, L., Wu, L. (2018), *Video copy detection based on deep cnn features and graph-based sequence matching*, Wireless Personal Communications 103(1), 401-416.
- [23] Zhao, G., Zhang, B., Zhang, M., Li, Y., Liu, J., Wen, J. (2022), *Star-gnn: spatial-temporal video representation for content-based retrieval*, In: ICME., 01-06.

## FEATURE ENGINEERING WITH CNN MODELS FOR PARTIAL VIDEO COPY DETECTION PROBLEM

Le Van Hao, Le Dinh Nghiep, Nguyen Van Cuong

### ABSTRACT

2D convolutional neural networks are the key component in partial video copy detection systems. They play a crucial role in video retrieval and matching tasks within a large database. However, the performance characteristics of these feature extraction methods have not thoroughly discussed in literature. This paper presents two key contributions. First, we conduct the experiments on a large-scale dataset to demonstrate the generalization capability and clarify the performance characteristics of popular neural networks. Next, we propose a time-series model approach to highlight the advantages and limitations of image features extracted from neural networks in the partial video copy detection problem.

**Keywords:** CNN models, feature engineering, partial video copy detection.

\* Ngày nộp bài: 19/03/2025; Ngày gửi phản biện: 21/03/2025; Ngày duyệt đăng: 25/9/2025