*DOI: 10.15625/vap.2022.0251*

# PHƯƠNG PHÁP LẬP CHỈ MỤC VÀ TÌM KIẾM VIDEO THEO NỘI DUNG

**Trang Thanh Trí, Phạm Thế Phi, Đỗ Thanh Nghị**
Khoa CNTT-TT, Trường Đại học Cần Thơ
*{tritrang, ptphi, dtnghi}@ctu.edu.vn*

***TÓM TẮT:*** *Bài báo này, chúng tôi trình bày phương pháp lập chỉ mục và tìm kiếm video theo nội dung, kết quả mô hình sẽ đánh giá trên tập dữ liệu được chúng tôi thu thập từ các bản tin thời sự của Đài Truyền hình Việt Nam. Bài toán gồm 2 giai đoạn: một là trích xuất keyframe và hai là truy vấn video dựa vào đặc trưng ảnh. Trong đó, ở giai đoạn một, chúng tôi đề xuất mô hình NET19_SBD cho bài toán phân loại chuyển cảnh (SBD) trong video. Kết quả thực nghiệm cho thấy phương pháp NET19_SBD mang lại sự chính xác cao (93,67 % cho lớp chuyển cảnh tức thời - Abrupt, 74,37 % cho lớp chuyển cảnh có hiệu ứng - Gradual) trong quá trình trích xuất ảnh đại diện cho video. Đầu ra của quá trình nhận dạng chuyển cảnh là các keyframe đại diện cho từng phân đoạn trong video, các keyframe này sẽ được trích đặc trưng và lập chỉ mục để hỗ trợ cho quá trình tìm kiếm. Tiếp theo, trong giai đoạn truy vấn, chúng tôi đề xuất mô hình kết hợp đặc trưng cục bộ bất biến (SIFT-1M) và đặc trưng mạng nơron tích chập với độ chính xác trung bình mAP@10 là 79 %.*

***Từ khóa:*** *Shot boundary detection (SBD), Query-by-image, Video retrieval, SIFT, CNN.*

## I. GIỚI THIỆU

Tìm kiếm video dựa vào nội dung là một trong những thử thách trong lĩnh vực thị giác máy tính. "Dựa vào nội dung", trong ngữ cảnh này chính là dựa vào những phân tích về nội dung trong video. Những "nội dung" này là những đặc trưng về màu sắc, hình dáng hay kết cấu của các khung hình trong video. Hiện nay, có nhiều nghiên cứu liên quan đến việc thao tác trên nội dung dữ liệu đa phương tiện, điển hình như: đề tài truy hồi video từ Bộ sưu tập video về lịch sử của Đơn vị Lưu trữ Phát thanh & Truyền hình Đức do nhóm nghiên cứu Markus Mühling thực hiện vào năm 2017 [1]. Trong đó, họ cung cấp lý thuyết thuật toán về mô hình phân lớp các khái niệm trực quan, tìm kiếm tương đồng, nhận dạng người và ký tự nhằm phục vụ mô hình tìm kiếm. Một đề tài khác cũng được thực hiện vào năm 2017 là đề tài Truy hồi video ca nhạc (MV) của Sungeun Hong và các cộng sự [2]. Mô hình của họ nhận dữ liệu đầu vào là các MV, những MV này được tách làm hai kênh để xử lý riêng biệt (nhạc và ảnh). Đối với kênh nhạc thì họ trích các đặc trưng cấp thấp thông qua các bộ nhận dạng và trích xuất đặc trưng âm thanh, còn với kênh ảnh thì họ sử dụng mô hình CNN đã được huấn luyện để trích đặc trưng. Sau đó, 2 đặc trưng này sẽ được đưa vào mô hình VM-NET của họ để rút trích đặc trưng đại diện cho MV nhằm phục vụ cho việc truy hồi MV dựa vào nhạc hoặc video. Năm 2018, Giorgos Kordopatis-Zilos cũng cho ra mắt bài báo với nhan đề "Fine-grained Incident Video Retrieval" [3] , bài báo này cung cấp cách giải quyết vấn đề truy vấn nội dung video từ một video đầu vào. Mô hình của họ dựa vào các mô hình mạng nơron tích chập VGG16, ResNet152 hay InceptionV4 để trích xuất các đặc trưng của key frame từ video. Kết quả họ đạt được 0,681 ở độ đo mAP trong tập dữ liệu FIVR-200k. Đối với bài toán phát hiện chuyển cảnh trong video, hướng tiếp cận học sâu đã đem lại kết quả rất khả quan so với các nghiên cứu trước đó. Điển hình là phương pháp [4] được đề xuất bởi Hassanien, phương pháp này nhận input đầu vào là chuỗi 16 khung hình, backbone là kiến trúc mô hình tác giả đề xuất C3D, output đầu ra được huấn luyện trên bộ phân lớp SVM. Ngoài ra, còn có các phương pháp [5] [6] của nhóm tác giả Tomáš Souček và cộng sự, phương pháp này sử dụng kiến trúc TransNet đạt kết quả tốt trên tập RAI, ClipShots, BBC tương ứng 93,9; 77,9; 96,2 ở độ đo F1 cũng như đạt được tốc độ tính toán ở thời gian thực. Hầu hết trong các nghiên cứu về truy hồi video, họ chỉ tập trung vào quá trình tìm kiếm mà bỏ qua bài toán phát hiện và nhận dạng chuyển cảnh trong video. Vì vậy, bài báo này sẽ thực hiện một nghiên cứu đầy đủ cho vấn đề tìm kiếm video theo nội dung. Nội dung bài báo có hai vấn đề giải quyết. Một là phát hiện và nhận dạng các chuyển cảnh trong video. Hai là giải quyết nhu cầu tìm kiếm video chứa các khung hình có sự tương đồng cao với ảnh truy vấn đầu vào bằng cách sử dụng các đặc trưng cơ bản như: đặc trưng cục bộ bất biến SIFT (Scale Invariant Feature Transform), đặc trưng mô tả toàn cục GIST, đặc trưng lược đồ hướng các gradient (Histogram of Oriented Gradients) và đặc trưng lược đồ màu (Color Histogram). Đồng thời, chúng tôi cũng sử dụng các đặc trưng học sâu bằng cách trích ra từ các mô hình mạng nơron tích chập VGG16 [4], VGG19 [4], ResNet50 [5], InceptionV3 [6] và Dense201 [7], các mô hình đã được huấn luyện nhận dạng 1000 đối tượng trong tập ImageNet. Thêm nữa, chúng tôi cũng đề xuất một phương pháp kết hợp cả đặc trưng cấp thấp và học sâu lại với nhau để gia tăng hiệu quả của quá trình tìm kiếm video.

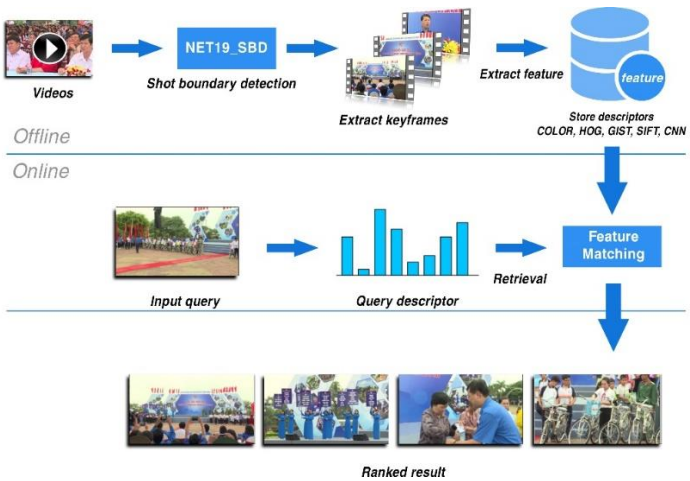
## II. HỆ THỐNG TÌM KIẾM VIDEO THEO NỘI DUNG

Hệ thống tìm kiếm video theo nội dung (Content-based Video Retrieval - CBVR) được chia thành hai giai đoạn chính theo sơ đồ ở Hình 1.

Giai đoạn 1 (offline) được xem là giai đoạn tiền xử lý của hệ thống. Ở giai đoạn này, tất cả video trong tập dữ liệu sẽ được tách từng khung hình đại diện (key frame) cho từng shot hình trong video thay vì phải trích xuất tất cả khung hình của video để lưu trữ. Để làm được điều này, hệ thống cần phải nhận dạng, phát hiện được các chuyển cảnh giữa các shot hình trong video. Bằng phương pháp mạng nơron tích chập, nghiên cứu đã xây dựng thành công mô hình phát hiện chuyển cảnh với tên gọi là NET19_SBD (NET19: là mô hình mạng với 19 lớp, SBD: Shot boundary

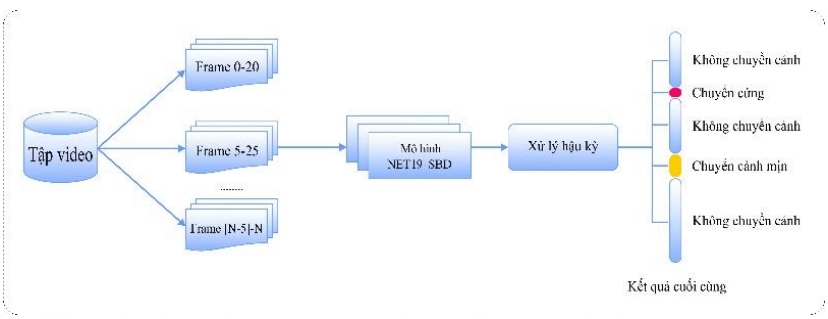
detection - Phát hiện và nhận dạng chuyển cảnh). Vì vậy, các video trong hệ thống sẽ được đưa vào bộ phân lớp NET19_SBD để phát hiện các khu vực có chuyển cảnh trong video và từ đó trích ra các khung hình đại diện cho các shot hình trong video. Kết thúc quá trình trích xuất khung hình cho video, hệ thống sẽ có được tập dữ liệu hình ảnh ứng với tập video trong cơ sở dữ liệu. Tiếp theo, trong giai đoạn offline này, từ tập dữ liệu hình ảnh đại diện sẽ được lần lượt trích các đặc trưng ảnh bằng các phương pháp đã được đề xuất ở trên. Kết quả của bước này là tập các véctơ đặc trưng ảnh và được lưu trữ vào cơ sở dữ liệu nhằm phục vụ cho quá trình truy vấn ở giai đoạn online. Các đặc trưng được sử dụng trong hệ thống này được chia làm hai nhóm. Nhóm đặc trưng cấp thấp (low-level) bao gồm: COLOR, HOG, GIST và SIFT. Ở nhóm đặc trưng học sâu bao gồm các đặc trưng được trích ra từ các mô hình CNN đã được huấn luyện trên tập ImageNet là: VGG16, VGG19, ResNet50, InceptionV3 và DenseNet201.

Trong giai đoạn 2 (online), ảnh truy vấn sẽ được trích các đặc trưng theo một cách tương tự ở giai đoạn 1 (offline). Kết quả sẽ được véctơ đại diện cho ảnh truy vấn. Sau đó, hệ thống sẽ tính toán độ tương đồng của ảnh đầu vào với ma trận véctơ đặc trưng của tập key frame trong cơ sở dữ liệu và xếp hạng trả về cho người dùng theo thứ tự đồng tương đồng cao nhất sẽ được nằm vị trí thứ nhất trong chuỗi kết quả trả về.




**Hình 1.** Sơ đồ tổng thể hệ thống

### *A.* Phát hiện và nhận dạng chuyển cảnh




**Hình 2**. Mô hình phát hiện và nhận dạng chuyển cảnh

Chúng tôi đề xuất một kỹ thuật tự động nhận dạng và phân loại chuyển cảnh trong video với tên gọi NET19_SBD. Một video sẽ được chia thành nhiều đoạn, mỗi đoạn có độ dài 20 khung hình. Các đoạn này sẽ được đưa vào mô hình và phân thành một trong ba lớp: chuyển cảnh không hiệu ứng, chuyển cảnh có hiệu ứng hoặc không chuyển cảnh. Chúng tôi sử dụng kiến trúc mạng NET19_SBD (tương tự kiến trúc VGG19), kiến trúc này nhận dữ liệu đầu vào là một ma trận kích thước (224,224,10) đại diện cho một đoạn video với độ dài 20 frame. Trong đó, phân đoạn này sẽ được chuẩn hóa về kích thước 224×224, chuyển kênh màu về không gian xám và lưu trữ 10 khung hình liên tiếp cách nhau bởi 1 frame. Kết quả đầu ra của mô hình NET19_SBD sẽ được tiếp tục xử lý hậu kỳ, bằng cách nhóm các đoạn video liên tiếp nhau có cùng nhãn chuyển cảnh. Chúng tôi huấn luyện mô hình nhận dạng NET19_SBD bằng tập dữ liệu cân bằng giữa các lớp chuyển cảnh không hiệu ứng, có hiệu ứng và không chuyển cảnh. Tập dữ liệu được chia ra 60 % cho việc huấn luyện, 20 % cho tập điều chỉnh tham số và 20 % cho tập kiểm nghiệm mô hình. Trong quá trình huấn luyện, chúng tôi tối ưu hàm lỗi bằng giải thuật Adam.

### *B.* Xây dựng tập dữ liệu đặc trưng ảnh

Chúng tôi sẽ sử dụng kết quả giai đoạn 1 để xây dựng tập ma trận đặc trưng ảnh. Với tập dữ liệu đầu vào, mỗi ảnh keyframe được chuẩn hóa về các kích thước tương ứng với từng mô hình (224×224, 299×299, 512×512). Sau đó, mỗi ảnh sẽ được trích chọn các đặc trưng về lược đồ màu (COLOR), lược đồ phân bố cạnh (HOG), đặc trưng toàn cục

(GIST), đặc trưng bất biến không đổi (SIFT) và các dạng đặc trưng học sâu bằng các phương pháp trích đặc trưng như VGG, Resnet, DenseNet. Chúng tôi xây dựng đầy đủ các đặc trưng từ đơn giản đến phức tạp, từ toàn cục đến cục bộ nhằm tìm ra một phương pháp tối ưu nhất cho việc phối hợp một cách linh hoạt các đặc trưng lại với nhau cho bài toán tìm kiếm video dựa vào nội dung.

***1.* Trích đặc trưng lược đồ màu**

Mỗi ảnh đại diện sẽ được xử lý bằng cách rời rạc hóa từng điểm màu sắc trong ảnh và phân vào 16 ngăn ở mỗi kênh màu (R,G,B). Giá trị mỗi điểm ảnh trong ngưỡng từ 0 đến 255 và được đưa vào ngăn thứ "value div 16". Kết quả sẽ được 1 ma trận kích thước (16,16,16). Khi kết hợp lại ta được một véctơ màu sắc đại diện cho ảnh đó với kích thước: 16×16×16 = 4096 chiều. Như vậy, kết quả cuối cùng sau khi trích xuất đặc trưng lược đồ màu cho tập dữ liệu ta được một ma trận có kích thước là (12990, 4096).

***2.* Trích đặc trưng lược đồ hướng gradient**

Sau khi trích đặc trưng lược đồ màu, tiếp tục trích đặc trưng HOG. Từng ảnh trong tập keyframe sẽ được trích đặc trưng theo cách bước sau: Ảnh đầu vào sẽ được tiền xử lý và điều chỉnh về kích thước 64×64. Sau đó, chuẩn hóa Gamma và Color cho ảnh. Tính Gradients: Ảnh sẽ được lấy gradient theo phương ngang x ($D_x$) và phương dọc y ($D_y$). Kết quả của ảnh X-Gradient gọi là $I_x$ và của Y-Gradient gọi là $I_y$. Bằng cách đơn giản là lọc ảnh với 2 kernel [-1, 0, 1] và $[-1, 0, 1]^T$. Tính lược đồ Gradients trong cell 8×8: ảnh đầu vào sẽ được chia thành các cell có kích thước 8×8 và tính lược đồ trên các cell đó, mỗi lược đồ có 9 bins. Tính véctơ đặc trưng: với mỗi cửa sổ trượt có kích thước 16×16 và số bước trượt là 8 thì số lần duyệt qua toàn ảnh kích thước 64×64 sẽ là: $\left[\frac{(64-16)}{8}+1\right] \times \left[\frac{(64-16)}{8}+1\right] = 7 \times 7 = 49$. Do block 16×16 được đại diện bởi 4 lược đồ, mỗi lược đồ 9 bins vì vậy mỗi block được đại diện bởi 1 véctơ: 4×9 = 36 chiều. Cuối cùng, kích thước đặc trưng HOG mỗi ảnh sẽ là: 7×7×4×9 = 1764 chiều. Kết quả sau khi trích đặc trưng HOG cho toàn tập dữ liệu sẽ được một ma trận có kích thước là (12990,1764).

***3.* Trích đặc trưng mô tả toàn cục**

Sau khi trích đặc trưng lược đồ hướng, chúng tôi tiếp tục trích đặc trưng GIST. Từng ảnh trong tập keyframe sẽ được trích đặc trưng theo cách bước sau:

- Mỗi ảnh trong tập keyframe sẽ được chuẩn hóa về kích thước 512×512 và được chuyển thành ảnh xám.
- Áp dụng phép biến đổi Fourier lên ảnh để đưa ảnh về miền tần số.
- Ứng với mỗi ảnh ở miền tần số sẽ được áp 32 bộ lọc Gabor lên ảnh. Bộ lọc Gabor được tạo ở 4 scale và 8 hướng (4×8=32 bộ lọc).
- Kết quả của mỗi bộ lọc sẽ được áp dụng phép biến đổi Fourier ngược để chuyển ảnh về miền không gian.
- Sau đó, các ảnh này sẽ được đưa về khung lưới 4×4 (16 ô vuông) và trích đặc trưng ứng với từng vùng trong ảnh.
- Như vậy, mỗi ảnh đầu vào sau khi trích đặc trưng GIST sẽ được 1 véctơ có kích thước là: (8+8+8+8) ×16 = 512 chiều.

Cuối cùng, sau khi trích đặc trưng GIST cho toàn tập dữ liệu sẽ được một ma trận có kích thước là (12990, 512).

***4.* Trích đặc trưng cục bộ bất biến**

Sau khi trích đặc trưng mô tả toàn cục GIST, chúng tôi tiếp tục trích đặc SIFT. Từng ảnh trong tập keyframe sẽ được trích đặc trưng theo cách bước: mỗi ảnh trong tập key frame sẽ được chuẩn hóa về kích thước có độ rộng là 299x168 và chuyển đổi thành ảnh xám. Các điểm đặc trưng trong ảnh sẽ được tính bằng cách áp dụng giải thuật phát hiện đặc trưng cục bộ lên ảnh như Harris-Affine, Hessian-Affine. Những điểm này có thể là cực trị cục bộ của phép toán DoG hoặc là cực đại của phép toán LoG. Sau đó, các vùng xung quanh các điểm đặc trưng này được xác định và mô tả bằng véctơ cục bộ. Mỗi một véctơ mô tả là một ma trận 4×4 các tổ chức đồ, mỗi tổ chức đồ có 8 hướng. Vì vậy, một điểm đặc trưng cục bộ bất biến được biểu diễn bằng một véctơ có kích thước là: 4×4×8=128 chiều. Với mỗi ảnh đại diện, sẽ thu được một ma trận số với số hàng là số điểm SIFT đã tìm được và số cột là 128 (chính là số chiều của đặc trưng bất biến). Như vậy, sau khi thu thập tất cả đặc trưng SIFT từ tập dữ liệu ta được một ma trận lớn, với số hàng là tất cả điểm cục bộ bất biến từ tất cả ảnh.

Kết quả của quá trình trích đặc trưng SIFT cho tập dữ liệu key frame sẽ được một ma trận lớn có kích thước là (4747672, 128).

***5.* Xây dựng từ điển từ trực quan, mô hình túi đặc trưng ảnh**

Kết quả của phần trước là ma trận đặc trưng cục bộ bất biến chính là tiền đề để xây dựng mô hình túi từ trực quan BoVW. Trong nghiên cứu này, chúng tôi xây dựng 3 bộ từ điển với kích thước khác nhau (nhỏ - 20k, vừa 200k và lớn 1.000.000 từ trực quan) bằng giải thuật gom cụm K-Means. Kết quả của giải thuật gom cụm chính là k tâm dữ liệu đại diện cho k cụm đặc trưng bất biến trong ma trận được xây dựng ở phần trước. Như vậy, chúng tôi đã có được 3 bộ từ điển khác nhau. Một ảnh trong tập keyframe sẽ được lượng hóa thành 3 véctơ tương ứng với 3 bộ từ điển đã được xây dựng. Để lượng hóa một ảnh vào trong bộ từ điển, chúng tôi tiến hành thực hiện các bước nhau sau:

**Bước 1**: Trích các đặc trưng cục bộ của ảnh.

**Bước 2**: Sử dụng mô hình K-means đã được xây ở trên để gán tâm dữ liệu cho từng đặc trưng đã tìm được ở bước 1.

**Bước 3**: Xây dựng lược đồ đặc trưng bằng cách đếm số lần xuất hiện các tâm tương ứng của kết quả ở bước 2. Mỗi lược đồ có kích thước (20k, 200k và 1M chiều).

Nhưng chúng tôi không lưu các véctơ đã được lượng hóa vào cơ sở dữ liệu vì sẽ rất tốn bộ nhớ để lưu trữ các véctơ có số chiều lớn như 200k và 1M chiều. Thay vào đó, chúng tôi sẽ lưu dữ liệu ở bước 2, các tâm của các đặc trưng cục bộ của ảnh theo cách lưu chỉ mục nghịch đảo (inverted index). Điều này vừa giúp chúng tôi tiết kiệm không gian lưu trữ và đẩy nhanh quá trình tính toán, lọc dữ liệu sau này. Cuối cùng, ở giai đoạn này, chúng tôi đã xây dựng được 3 tập chỉ mục nghịch đảo của 12.990 ảnh đại diện ứng với 3 bộ từ điển đặc trưng.

### *6.* Trích đặc trưng học sâu bằng phương pháp mạng nơron tích chập

Sau các giai đoạn trích đặc trưng cấp thấp, chúng tôi tiến hành trích tiếp các đặc trưng học sâu bằng cách sử dụng các mô hình mạng nơron tích chập đã được huấn luyện trên tập ImageNet như VGG16, VGG19, ResNet50, InceptionV3 và DenseNet201. Tất cả đặc trưng này đều được trích bằng các thao tác tương tự như: lấy các mô hình mạng này và cắt bỏ layer cuối cùng (layer dự đoán lớp). Sau đó, đưa trực tiếp các ảnh vào các mô hình này để trích ra các feature map tương ứng. Kết quả cuối cùng:

**VGG16**: Sẽ được ma trận đặc trưng có kích thước (12990, 4096).

**VGG19**: Sẽ được ma trận đặc trưng có kích thước (12990, 4096).

**ResNet50**: Sẽ được ma trận đặc trưng có kích thước (12990, 1000).

**InceptionV3**: Sẽ được ma trận đặc trưng có kích thước (12990, 2048).

**DenseNet201**: Sẽ được ma trận đặc trưng có kích thước (12990, 94080)

### *7.* Truy vấn video dựa vào nội dung

Đến đây, giai đoạn offline đã hoàn thành. Kết quả, hệ thống đã có tất cả ma trận đặc trưng của tập ảnh đại diện video. Tiếp theo là giai đoạn online, nhiệm vụ của giai đoạn này là nhận một ảnh đầu vào và tính toán trả về tập keyframe có độ tương đồng cao cho người dùng. Ảnh truy vấn sẽ được đưa vào hệ thống và trích các đặc trưng tương tự như cách đã làm ở giai đoạn offline. Ứng với từng đặc trưng khác nhau mà hệ thống tính toán trên từng ma trận đặc trưng ấy. Cuối cùng, hệ thống trả về kết quả đã được sắp xếp độ tương đồng theo thứ tự giảm dần cho người dùng.

## III. KẾT QUẢ THỰC NGHIỆM

Chúng tôi trình bày kết quả thực nghiệm hai bài toán, một là bài toán phát hiện và phân lớp chuyển cảnh dùng cho việc tách khung hình đại diện trong video nhằm phục vụ cho công việc lập chỉ mục (indexing video), hai là bài toán truy hồi video bằng hình ảnh (retrieval video by image). Toàn bộ nghiên cứu được cài đặt bằng ngôn ngữ lập trình Python cùng với việc sử dụng các thư viện hỗ trợ như Keras cho việc xây dựng và huấn luyện kiến trúc mạng nơron tích chập, Tensorflow cho việc tính toán giải thuật gom cụm K-Means bằng GPU, thư viện OpenCV để trích đặc trưng SIFT, COLOR, HOG và thư viện Scikit-learn để xây dựng mô hình túi đặc trưng từ trực quan. Ngoài ra, nghiên cứu còn sử dụng thư viện FFmpeg để xây dựng tập dữ liệu tăng cường cho mô hình nhận dạng chuyển cảnh trong video.

### *A.* Chuẩn bị dữ liệu

Có 4 tập dữ liệu được sử dụng: một là tập TRECVID 2001 (1 phần dùng cho huấn luyện, 1 phần dùng cho kiểm nghiệm và so sánh với các mô hình khác), hai là tập nhận dạng đối tượng trong video của ImageNet (tập này dùng để sinh ra dữ liệu tăng cường cho việc huấn luyện mô hình Net19_SBD), ba là tập video từ bản tin thời sự của Đài Truyền hình Việt Nam VTV (là tập kiểm nghiệm chính trong nghiên cứu này), bốn là tập dữ liệu được tạo ra nhằm tăng cường dữ liệu cho việc huấn luyện mô hình CNN nhận dạng chuyển cảnh.

1) TRECVID 2001: Đây là tập dữ liệu được sử dụng trong chuỗi hội thảo TREC được tài trợ bởi Viện Tiêu chuẩn và Công nghệ Quốc gia (NIST) với sự hỗ trợ bổ sung từ các cơ quan chính phủ khác của Hoa Kỳ.

2) Tập nhận dạng đối tượng trong video của ImageNet: Tập dữ liệu này được sử dụng trong thử thách nhận dạng đối tượng trong video của ImageNet. Trong nghiên cứu này, sẽ sử dụng tập dữ liệu này nhằm sinh ra tập dữ liệu tăng cường cho việc huấn luyện mô hình phân lớp chuyển cảnh trong video. Tập dữ liệu này bao gồm 7314 tập tin video với tổng dung lượng là 28,8 GB và thời lượng là 25 tiếng.

3) Tập video từ các bản tin thời sự VTV: Đây là tập dữ liệu chính trong nghiên cứu này, được sử dụng để đánh giá mô hình đề xuất. Tập dữ liệu này bao gồm 618 tập tin, tổng dung lượng được nén chuẩn H264 là 3,84 GB với thời lượng 13 giờ 26 phút. Trong đó, 100 tập tin được chọn ngẫu nhiên để đánh giá mô hình phân lớp chuyển cảnh. Đồng thời, lấy ngẫu nhiên 100 ảnh đại diện ứng với 100 video trong tập kiểm nghiệm này để xây dựng tập truy hồi video.

4) Tập dữ liệu tăng cường: Là tập dữ liệu được sinh ra từ tập ImageNet nhằm để tăng cường dữ liệu cho mô hình CNN. Tập dữ liệu này bao gồm 14.000 tập tin cho mỗi lớp chuyển cảnh với thời lượng 1 giây cho mỗi tập tin.

### *B.* Cài đặt chương trình

Chúng tôi đã tiến hành cài đặt tìm kiếm video theo nội dung bằng ngôn ngữ lập trình Python. Các thuật toán VGG, ResNet, Dense được cài đặt bằng python và sử dụng thư viện Keras [16] với backend là Tensorflow [17]. Tất cả kết quả nghiên cứu đều được thực nghiệm trên cùng hệ thống máy tính có kiến trúc như sau: Chúng tôi cài đặt

- CPU: Intel Xeon E5-2667 @2.90GHz (6 cores);
- GPU: 2x Nvidia GTX 1080Ti, 11GB GDDR5;
- RAM: 32GB DDR3;
- Operating system: Windows 10;
- Library: python 3.7.6, keras 2.2.4, tensorflow 1.15.0.

### *C.* Kết quả thực nghiệm

Nghiên cứu tiến hành thực nghiệm trên hai bài toán: phân lớp chuyển cảnh và tìm kiếm nội dung video bằng tất cả phương pháp đã đề xuất và sử dụng các tập dữ liệu đã thu thập được để huấn luyện và kiểm nghiệm. Trong nghiên cứu này, sẽ sử dụng độ đo Precision, Recall và F1 để đánh giá mô hình phân lớp chuyển cảnh và độ đo mAP để đánh giá mô hình tìm kiếm nội dung video.

*1.* Phát hiện và nhận dạng chuyển cảnh video

Chúng tôi tiến hành huấn luyện mô hình phân lớp trên dữ liệu tăng cường được sinh ra từ tập dữ liệu chuẩn "ImageNet object detection video" kết hợp tập huấn luyện của TREC2001. Mô hình được huấn luyện trên tập 102.400 mẫu, 33.225 mẫu trong tập điều chỉnh tham số, kiểm nghiệm trên tập 33.553 mẫu và kết quả đạt được 97,5 %. Chúng tôi đánh giá kết quả bằng cách sử dụng các độ đo: độ chính xác (P), độ bảo phủ (R) và độ đo F1. Đồng thời, chúng tôi sử dụng chỉ số đánh giá chuẩn TRECVID để đánh giá một đoạn video là chuyển cảnh hiệu ứng khi nó chứa ít nhất một khung hình chuyển cảnh. Trong quá trình so sánh, chúng tôi sẽ tô đậm các kết quả có độ chính xác cao nhất. Nghiên cứu này sẽ so sánh kết quả với các kỹ thuật mới nhất (DeepSBD [8], Priya et al. [9]). Trong các phương pháp trên đã cho thấy điểm mạnh về độ chính xác trong việc nhận dạng và còn những hạn chế về tốc độ xử lý. DeepSBD [8] là phương pháp đạt được tốc độ cao nhất so với các phương pháp còn lại. Kết quả chúng tôi trình bày sẽ cho thấy việc tối ưu cả về độ chính xác lẫn tốc độ nhận dạng trên đồng thời cả tập TREC2001 và Thời sự VTV.

**Bảng 1.** So sánh kết quả thực nghiệm độ đo F1 trên tập TREC2001

| Method | D1-anni5 | D4-anni10 | D5-NAD57 | D6-NAD58 |
|---|---|---|---|---|
| Priya et al. [9] | | | | |
| Abrupt | 0.85 | 0.897 | 0.945 | 0.945 |
| Gradual | 0.938 | 0.822 | 0.809 | 0.885 |
| DeepSBD [8] | | | | |
| Abrupt | 0.818 | 0.918 | **0.957** | 0.904 |
| Gradual | **0.945** | **0.855** | **0.917** | 0.914 |
| Net19_SBD | | | | |
| Abrupt | **0.907** | **0.919** | 0.926 | **0.95** |
| Gradual | 0.88 | 0.688 | 0.875 | **0.921** |

*2.* Tìm kiếm video theo nội dung ảnh

Trong phần này, nghiên cứu sẽ trình bày kết quả tìm kiếm video dựa trên nội dung ảnh truy vấn bằng các phương pháp khác nhau đối với tập dữ liệu thực của Đài Truyền hình Việt Nam. Nghiên cứu sẽ so sánh các kết quả tìm kiếm dựa vào việc sử dụng các đặc trưng cấp thấp như: COLOR, HOG, GIST, SIFT và đặc trưng học sâu đã được huấn luyện từ các mô hình VGG16, VGG19, ResNet50, InceptionV3, DenseNet201 và kết hợp giữa hai đặc trưng cấp thấp và cao. Tiêu chí đánh giá mô hình là dựa vào độ đo mAP trên TOP 10, 20, 30, 40, 50 kết quả trả về, cũng như thời gian thực thi truy vấn trên mỗi phương pháp. Để kiểm nghiệm mô hình, nghiên cứu tiến hành xây dựng tập kiểm tra bằng cách chọn ngẫu nhiên 100 video trong tập dữ liệu. Sau đó, tiến hành lấy ngẫu nhiên các keyframe trong tập key frame được xây dựng từ kết quả của bài toán nhận dạng chuyển cảnh trong video, kết quả sẽ được 100 ảnh truy vấn ngẫu nhiên tương ứng với 100 video ngẫu nhiên được chọn. Tiến hành thực nghiệm, giai đoạn offline sẽ trích tất cả đặc trưng từ các keyframe trong tập dữ liệu (12.990 key frame tương ứng 618 video), kết quả giai đoạn này là 8 ma trận đặc trưng tương ứng với 8 đặc trưng đã nêu ở trên. Riêng đặc trưng cục bộ bất biến SIFT được xây dựng làm 3 tập chỉ mục riêng biệt tương ứng với 3 kích thước từ điển khác nhau từ thấp đến cao (20.000 đặc trưng: thấp, 200.000 đặc trưng: trung bình, 1.000.000 đặc trưng: cao). Giai đoạn online, ảnh truy vấn sẽ được trích đặc trưng tương ứng với từng phương pháp. Véctơ đặc trưng ảnh truy vấn sẽ được tính độ tương đồng với tập dữ liệu và trả về kết quả trong TOP 10, 20, 30, 40, 50 để so sánh mAP giữa các phương pháp. Kết quả được trình bày chi tiết trong Bảng 2.

**Bảng 2.** Kết quả thực nghiệm bài toán truy hồi video bằng đặc trưng ảnh

| | Method | DIM | Index | TF-iDF | mAP @10 | mAP @20 | mAP @30 | mAP @40 | mAP @50 | Time (s) |
|---|---|---|---|---|---|---|---|---|---|---|
| Low-level feature | COLOR | 4096 | - | - | 42,56% | 39,93% | 37,21% | 36,36% | 35,81% | 0,562 |
| | HOG | 1764 | - | - | 16,31% | 15,52% | 15,10% | 14,65% | 13,84% | **0,281** |
| | GIST | 512 | - | - | 35,87% | 35,56% | 34,70% | 33,14% | 32,46% | 1,65 |
| | SIFT-based Bag of feature | 20K | inv | - | 37,88% | 36,16% | 33,66% | 32,64% | 30,77% | 4,96 |
| | | 20K | inv | TRUE | 38,78% | 36,78% | 34,94% | 33,10% | 31,46% | 4,98 |
| | | 200K | inv | - | 42,01% | 40,04% | 38,35% | 37,78% | 37,18% | 6,43 |
| | | 200K | inv | TRUE | 42,54% | 39,66% | 38,11% | 36,96% | 36,75% | 6,42 |
| | | 1M | inv | - | **45,58%** | 44,81% | 42,82% | **41,89%** | **41,80%** | 11,87 |
| | | 1M | inv | TRUE | 45,35% | **44,86%** | **43,83%** | 41,66% | 40,86% | 12,17 |
| Pre-trained CNN-based (High-level feature) | VGG16 | 4096 | - | - | **76,05%** | **69,87%** | **67,11%** | **64,54%** | **62,73%** | 0,578 |
| | VGG19 | 4096 | - | - | 74,21% | 67,50% | 64,14% | 61,70% | 59,93% | 0,578 |
| | ResNet50 | 1000 | - | - | 33,64% | 30,51% | 27,23% | 25,94% | 24,34% | 0,203 |
| | InceptionV3 | 2048 | - | - | 70,62% | 64,21% | 60,52% | 57,72% | 54,79% | 0,343 |
| | DenseNet201 | 94080 | - | - | 73,40% | 67,99% | 63,78% | 60,72% | 58,79% | 19,87 |
| Combine | SIFT-1M, VGG16 | 1M, 4096 | inv-index | TRUE | **78,99%** | **72,51%** | **69,38%** | **67,44%** | **65,39%** | 13,61 |

Trong nhóm các đặc trưng cấp thấp như: COLOR, HOG, GIST, SIFT thì đặc trưng HOG cho kết quả không tốt nhất, đặc trưng này không phù hợp với tập dữ liệu video ở dạng thời sự. Còn đặc trưng GIST, có độ chính xác vào khoảng 32~35 %, đặc trưng này với số chiều dữ liệu là thấp nhất (512 chiều) nhưng thời gian thực hiện một lệnh truy vấn tốn đến 1,62 giây (do quá trình trích đặc trưng GIST tốn khá nhiều chi phí tính toán và chỉ dựa vào CPU). Đặc trưng lược đồ màu mang lại hiệu quả khá cao cho bài toán này, với độ chính xác trung bình rơi vào khoảng 35~42 % và chỉ tốn 0,562 giây cho 1 lần truy vấn. Điều này cho thấy, trong tập dữ liệu kiểm nghiệm các keyframe có sự biến động về màu sắc khá khác nhau. Riêng mô hình túi đặc trưng ảnh BoVW (đặc trưng SIFT) một lần nữa lại được chứng minh về sự mạnh mẽ cho bài toán tìm kiếm nội dung ảnh. Kết quả cho thấy, mô hình BoVW dễ dàng tăng độ chính xác bằng cách tăng số lượng đặc trưng đại diện (cluster). Trong bài toán này, mô hình BoVW với số lượng đặc trưng 20.000 chỉ đạt 31~38 %, khi tăng số lượng cluster lên 200.000 thì đạt 36~42 %, và khi tăng lên đến 1.000.000 cluster thì độ chính xác trung bình lên cao nhất là 45,58 %. Cũng trong mô hình BoVW, các véctơ được tính trọng số Tf-iDF cũng cải thiện được rất ít về độ chính xác. Nhược điểm lớn nhất của mô hình BoVW chính là chi phí tính toán quá cao, một ảnh truy vấn tốn tới 11-12 giây cho mô hình BoVW-1M và ~5 giây cho BoVW-20k.

Trong nhóm đặc trưng học sâu, ngoài đặc trưng được trích từ mô hình ResNet50 cho độ chính xác khá thấp 24~33 % thì các đặc trưng từ mô hình khác (VGG16, VGG19, InceptionV3, DenseNet201) đều cho kết quả rất tốt 70~76 % tại TOP 10. Nhóm đặc trưng này vượt trội hơn hẳn so với nhóm đặc trưng cấp thấp như COLOR, HOG, GIST, SIFT về độ chính xác và cả thời gian thực thi, vì phần lớn tính toán được xử lý trên GPU.

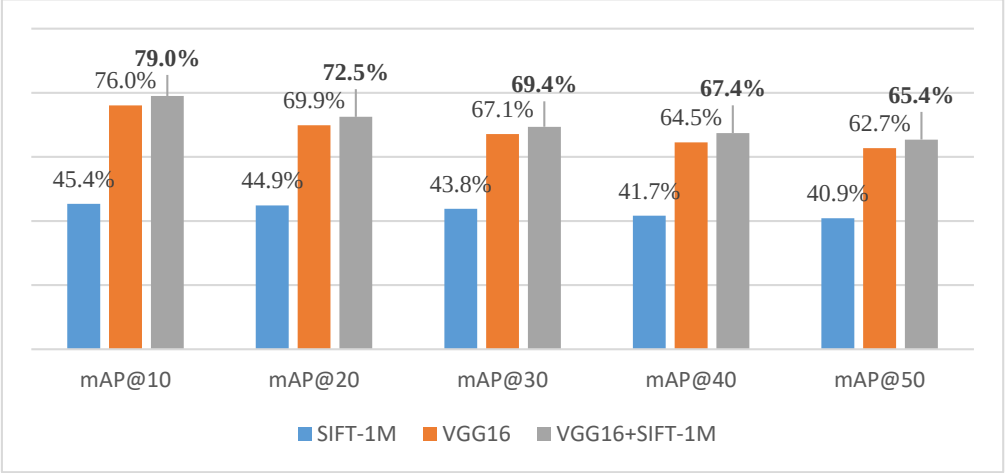
Từ kết quả trên, nghiên cứu tiếp tục chọn ra ứng viên từ hai nhóm đặc trưng này, để kết hợp lại với nhau. Trong nhóm đặc trưng cấp thấp, nghiên cứu chọn mô hình BoVW với số lượng đặc trưng là 1.000.000 và có sử dụng trọng số TF-iDF, trong nhóm đặc trưng học sâu thì nghiên cứu chọn phương pháp sử dụng đặc trưng học sâu được trích ra từ mô hình VGG16. Kết quả tính toán cuối cùng của việc kết hợp hai đặc trưng sẽ được tính theo công thức sau:

$$Result_{Final} = Alpha \times Result_{SIFT-1M} + (1 - Alpha) \times Result_{VGG16}$$

trong đó, Alpha là hệ số được tính cho kết quả trả về dựa trên đặc trưng SIFT-1M. Kết quả kết hợp hai đặc trưng này được trình bày cụ thể trong Bảng 3.

**Bảng 3.** Kết quả thực nghiệm kết hợp đặc trưng VGG16 và SIFT-1M theo tham số alpha

| Alpha | mAP@10 | mAP@20 | mAP@30 | mAP@40 | mAP@50 |
|---|---|---|---|---|---|
| 0.9 | 76,89% | 71,02% | 67,27% | 64,78% | 62,98% |
| 0.8 | 78,99% | **72,51%** | **69,38%** | **67,44%** | **65,39%** |
| 0.7 | **79,09%** | 72,21% | 69,19% | 67,16% | 65,00% |
| 0.6 | 78,78% | 71,16% | 68,22% | 66,24% | 64,21% |
| 0.5 | 77,96% | 70,76% | 67,25% | 65,32% | 63,24% |
| 0.4 | 77,60% | 70,79% | 67,88% | 65,05% | 62,98% |
| 0.3 | 77,27% | 70,76% | 67,92% | 65,32% | 63,30% |
| 0.2 | 76,38% | 70,29% | 67,46% | 64,89% | 62,98% |
| 0.1 | 75,88% | 69,76% | 66,99% | 64,46% | 62,58% |




**Hình 3.** Biểu đồ so sánh kết quả các loại đặc trưng thấp, cao và kết hợp

Từ các dữ liệu được trình bày ở Bảng 3 và 4, nghiên cứu cho thấy việc kết hợp hai kết quả trả về của đặc trưng SIFT và VGG16 theo hệ số alpha là 0,8 đã cải thiện được trung bình 3 % tại TOP 10, 20, 30, 40 và 50 (so với kết quả từ đặc trưng VGG16).

## IV. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Trong nghiên cứu, chúng tôi đã đề xuất hai phương pháp để giải quyết hai bài toán trong lĩnh vực tìm kiếm video. Một là bài toán phân lớp chuyển cảnh trong video, mục tiêu của bài toán này là giải quyết vấn đề video summary (tóm lược video bằng keyframe) nhằm để lập chỉ mục keyframe cho video để tiết kiệm chi phí tính toán thay vì phải tìm kiếm trên tất cả keyframe trong video. Bằng việc nghiên cứu các mô hình mạng nơron tích chập, nghiên cứu đã tìm ra mô hình NET19_SBD. Mô hình này chỉ với 75 triệu tham số đã đạt được kết quả rất khả quan, trung bình độ bao phủ của hai lớp Abrupt và Gradual là 97 % trên tập kiểm nghiệm và trung bình độ đo F1 của hai lớp này đạt 84 %. Thứ hai là bài toán tìm kiếm video bằng nội dung ảnh, nghiên cứu đề xuất phương pháp kết hợp giữa hai loại đặc trưng cấp thấp và học sâu. Ở đặc trưng học sâu, trong nghiên cứu này sử dụng đặc trưng được trích ra từ mô hình VGG16. Còn với đặc trưng cấp thấp, nghiên cứu đề xuất sử dụng phương pháp mô hình túi từ đặc trưng BoVW với số chiều từ điển là 1.000.000 có sử dụng TF-IDF để tính trọng số cho véctơ trong mô hình. Trọng số tốt nhất khi kết hợp 2 đặc trưng này với nhau là 0,2 cho VGG16 và 0,8 cho SIFT-1M. Kết quả đạt được từ phương pháp này là 79 % tại TOP 10 và 72,5 % tại TOP 20.

Về mặt thực tiễn: bài báo đã phân tích bài toán nhận dạng chuyển cảnh trong video và tìm kiếm video theo nội dung cho tập dữ liệu cụ thể là tập video các bản tin thời sự của Đài Truyền hình Việt Nam. Nghiên cứu đã tiến hành thực nghiệm nhiều phương pháp khác nhau để rút ra phương pháp phù hợp với tập dữ liệu này nhất. Đối với nhóm đặc trưng cấp thấp thì đặc trưng HOG cho kết quả thấp và không phù hợp với tập dữ liệu này nhất. Trong khi đó, đặc trưng SIFT với mô hình túi đặc trưng ảnh - 1.000.000 chiều mang lại kết quả tìm kiếm tốt nhất trong nhóm các đặc trưng cấp thấp. Còn với đặc trưng học sâu, thì đặc trưng được trích ra mô hình VGG16 đã được huấn luyện từ tập dữ liệu ImageNet mang lại độ chính xác trung bình tốt nhất. Cuối cùng, khi kết hợp hai loại đặc trưng SIFT-1M và đặc trưng VGG16 mang lại độ chính xác cao nhất vì đặc trưng được rút trích vừa mang được những tính chất của đặc trưng cấp thấp và cả đặc trưng học sâu. Ngoài ra, đóng góp của nghiên cứu này còn có hai tập dữ liệu lớn và đã được gán nhãn hoàn chỉnh cho hai bài toán phát hiện chuyển cảnh trong video và tìm kiếm video theo nội dung ảnh.

## TÀI LIỆU THAM KHẢO

[1] Markus Mühling, Manja Meister, Nikolaus Korfhage, Jörg Wehling, Angelika Hörth, Ralph Ewerth; and Bernd Freisleben, “Content-Based Video Retrieval in Historical Collections of the German Broadcasting Archive”, 19 February 2018.

[2] Sungeun Hong and Woobin Im and Hyun Seung Yang, “CBVMR: Content-Based Video-Music Retrieval Using Soft Intra-Modal Structure Constraint”, 2018.

[3] Kordopatis-Zilos, Giorgos and Papadopoulos, Symeon and Patras, Ioannis and Kompatsiaris, Ioannis, “FIVR: Fine-grained Incident Video Retrieval”, 2018.

[4] Karen Simonyan, Andrew Zisserman, “Very Deep Convolutional”, *ICLR*, 2015.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, “Deep Residual Learning for Image Recognition”, *ILSVRC*, 2015.

[6] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna, “Rethinking the Inception Architecture for Computer Vision”, 2015.

[7] Gao Huang, Zhuang Liu, Laurens van der Maaten, “Densely Connected Convolutional Networks”, *CVPR*, 2017.

[8] Ahmed Hassanien, Mohamed Elgharib, Ahmed Selim, Sung-Ho Bae, Mohamed Hefeeda, and Wojciech Matusik, “Large-scale, Fast and Accurate Shot Boundary Detection through Spatio-temporal Convolutional Neural Networks”, 2017.

[9] L. Priya and D. S, "Walsh hadamard transform kernel-based feature vector for shot boundary detection", *IEEE Transactions on Image Processing*, Vol. 23, p. 5187-5197, 2014.

[10] Tomáš Souček, Jaroslav Moravec, Jakub Lokoč, "TransNet: A deep network for fast detection of common shot transitions", 2019.

[11] Tomáš Souček and Jakub Lokoč, "TransNet V2: An effective deep network architecture for fast shot transition detection", 2020.

## CONTENT-BASED VIDEO INDEXING AND RETRIEVAL

**Trang Thanh Tri, Pham The Phi, Do Thanh Nghi**

***ABSTRACT:*** *In this paper, we present a very specific way of searching for video clips, which is based on the content of the video. We try to address two problems: one is keyframe extraction and the other is video query based on image features. In the first stage, we propose a method named NET19_SBD for detection and classification of video shots boundaries which is based on Convolutional Neural Network. Second, we also study the effectiveness of several low-level features such as Color Histogram, HOG, GIST, SIFT; high-level (semantic) features which are extracted from pretrained CNNs such as VGG16, VGG19, ResNet50, InceptionV3 and DenseNet201 and combine featurse in Content-based video retrieval. Finally, we perform our experiments on the data which are collected from Vietnam Television's news videos. Our approach "NET19_SBD" is high accuracy (93.67 % for sharp transition detection and 74.37 % for gradual transition). For video retrieval, the combination of SIFT-1M feature and high-level feature from VGG16 reach 79 % mAP@10.*