HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN THANH TÙNG

NGHIÊN CỨU THUẬT TOÁN PHÁT HIỆN ĐIỂM CẮT, GHÉP TRONG VIDEO

LUẬN VĂN THẠC SĨ KỸ THUẬT
(Theo định hướng ứng dụng)

HÀ NỘI - NĂM 2021

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

NGUYỄN THANH TÙNG

NGHIÊN CỨU THUẬT TOÁN PHÁT HIỆN ĐIỂM CẮT, GHÉP TRONG VIDEO

CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN

MÃ SỐ: 8.48.01.04

LUẬN VĂN THẠC SỸ KỸ THUẬT (HỆ THỐNG THÔNG TIN)

NGƯỜI HƯỚNG DẪN: PGS TS HÀ HẢI NAM

HÀ NỘI - NĂM 2021

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn về đề tài "Tìm hiểu về thuật toán phát hiện điểm cắt, ghép trong video" là công trình nghiên cứu cá nhân của tôi trong thời gian qua. Mọi số liệu sử dụng phân tích trong luận văn và kết quả nghiên cứu là do tôi tự tìm hiểu, phân tích một cách khách quan, trung thực, có nguồn gốc rõ ràn. Tôi xin chịu hoàn toàn trách nhiệm nếu có sự không trung thực trong thông tin sử dụng trong luận văn

LỜI CẢM ƠN

Trước hết em xin cảm ơn các thầy trong Ban giám hiệu, thầy cô trong Khoa Sau đại học cùng các giảng viên trong khoa Công nghệ thông tin I – Trường Học viện công nghệ bưu chính viễn thông đã tạo mọi điều kiện thuận lợi cho em trong quá trình học tập tại trường. Đặc biệt em xin chân thành cảm ơn sự hướng dẫn tận tình của thầy PGS.TS Hà Hải Nam - Phó Viện trưởng phụ trách Viện Công nghiệp phần mềm và Nội dung số Việt Nam đã tạo mọi điều kiện giúp đỡ em hoàn thành luận văn.

Mặc dù đã cố gắng hết sức cùng sự tận tâm của thầy giáo hướng dẫn xong do kiến thức còn hạn chế, nội dung nghiên cứu còn tương đối mới và khó với em nên luận văn không tránh khỏi những sai sót trong quá trình tiếp nhận kiến thức, nghiên cứu. Em rất mong chỉ dẫn của thầy cô và sự góp ý của bạn bè, đồng nghiệp để em có thể hoàn thiện luận văn của mình.

Cuối cùng em xin gửi lời cảm ơn đặc biệt nhất tới gia đình, bố, mẹ, những người động viên, khích lệ giúp em hoàn thành luận văn này.

Em xin chân thành cảm ơn!

Hà Nội, ngày tháng năm 2021

Người thực hiện

Nguyễn Thanh Tùng

MỤC LỤC

DANH MỤC CÁC THUẬT NGỮ, CHỮ VIẾT TẮT

Viết tắt	Tiếng Anh	Tiếng Việt
AWOB	Adjustable Width Object Boundary	Ranh giới đối tượng với chiều rộng có thể thay đổi được
AVC	Advanced Video Coding	Mã hóa video cao cấp
AMI	Advanced Metering Infrastructure	Hạ tầng đo đếm tiên tiến
AI	Artificial Intelligence	Trí tuệ nhân tạo
DCT	Discrete Cosine Transform	Biến đổi Cosine rời rạc
GMM	Gaussian Mixture Models	Mô hình hỗn hợp Gaussian
GOP	Group Of Picures	Nhóm các ảnh

DANH SÁCH BẢNG

DANH SÁCH HÌNH ẢNH

MỞ ĐẦU

1. Lý do chọn đề tài

Ngày nay, cùng với sự phát triển của khoa học và công nghệ, đặc biệt là ảnh hưởng của cuộc cách mạng công nghiệp 4.0 đã làm thay đổi mọi mặt của đời sống, xã hội; việc ứng dụng các công nghệ hiện đại như Trí tuệ nhân tạo (AI), Dữ liệu lớn (Big data), Dữ liệu nhanh (Fast data), Block chain… đã thúc đẩy sự phát triển của mọi lĩnh vực, từ kinh tế, văn hóa, truyền thông, khoa học kỹ thuật… cho đến công tác quản lý xã hội, đấu tranh phòng, chống tội phạm. Trong đó, sự ra đời và ứng dụng các phương tiện ghi âm, ghi hình trong công tác điều tra, phá án cũng như tố tụng ngày càng được triển khai sâu rộng, phổ biến trên thế giới; dữ liệu hình ảnh, video thu được từ các hiện trường vụ án đã trở thành một nguồn chứng cứ quan trọng, giúp cơ quan chức năng củng cố chứng cứ, chứng minh các hoạt động phạm tội. Tuy nhiên, bên cạnh những thuận lợi do sự phát triển của khoa học kỹ thuật hiện đại đem lại đó, nó cũng kéo theo nhiều ảnh hưởng tiêu cực trong đời sống; việc các video/hình ảnh giả mạo, chứa thông tin sai sự thật (*Deep-fakes*), các video/hình ảnh hiện trường bị chỉnh sửa, cắt ghép, bị các đối tượng phạm tội tác động làm sai lệch thông tin ngày càng phổ biến. Thế giới đã và đang phải đối mặt với nguy cơ thông tin sai sự thật, đặc biệt là qua các video giả mạo người nổi tiếng, lan tràn ngày càng nhiều trên Internet; các cơ quan chức năng thực thi pháp luật các nước đã phải đối mặt với vấn đề, thách thức trong việc phát hiện chỉnh sửa trong video chứng cứ từ lâu; tại nhiều quốc gia phát triển như Mỹ, Trung Quốc, Nga, Anh… nhiều công nghệ kỹ thuật đã được sử dụng để phát hiện việc các video/hình ảnh bị chỉnh sửa, giả mạo, qua đó phục vụ đắc lực cho lực lượng thực thi pháp luật nói chung và người dùng Internet nói riêng.

Tại Việt Nam, công tác giám định hình ảnh cũng được Viện Khoa học hình sự - Bộ Công an nghiên cứu, triển khai đạt được nhiều kết quả tích cực; tuy nhiên, do số lượng vụ án hàng năm ngày càng tăng, dữ liệu video thu được từ hiện trường các vụ án ngày càng lớn đã làm tăng cao nhu cầu phát hiện video giả mạo, bị chỉnh

sửa. Đáng chú ý, hiện nay công tác giám định video giả mạo cắt ghép chủ yếu được thực hiện hoàn toàn thủ công dựa trên quan sát trực tiếp video của các chuyên gia. Công việc này tốn rất nhiều thời gian và công sức đặc biệt khi các đoạn video thu từ camera có thời lượng lớn. Do đó, việc tự động hoá phát hiện video bị cắt ghép là nhu cầu cấp bách trong công tác điều tra, phá án. Nếu ứng dụng thành công các công nghệ, kỹ thuật hiện đại, hệ thống phát hiện video bị cắt ghép, giả mạo sẽ giúp giảm công sức của các chuyên gia và tăng hiệu quả xử lý công tác giám định kỹ thuật hình sự.

Với yêu cầu thực tiễn nêu trên, học viên đã chọn đề tài "*Nghiên cứu thuật toán phát hiện điểm cắt, ghép trong video*" với mục tiêu nghiên cứu một số giải pháp kỹ thuật phổ biến trên thế giới qua đó ứng dụng xây dựng hệ thống phần mềm giải quyết các bài toán thực tiễn.

2. Tổng quan về vấn đề nghiên cứu

Video đã trở thành một phần không thể thiếu trong giao tiếp hiện đại. Các trang web như YouTube và Facebook, các ứng dụng như Instagram và Twitter, cho phép người dùng ngay lập tức chia sẻ video với những người khác trên toàn thế giới. Tuy nhiên, việc chỉnh sửa video ngày càng trở nên dễ dàng hơn; trong đó, rất dễ dàng để một số người dùng tạo video được chỉnh sửa với ý đồ xấu. Kết quả là các video giả mạo và thông tin sai lệch được chia sẻ nhanh hơn trước khi chúng có thể được xác minh. Điều này đặt ra các câu hỏi về tính xác thực của nhiều video.

Gần đây, Deepfakes đã nổi lên như một mối đe dọa mới, thu hút sự chú ý của cả các nhà nghiên cứu và giới truyền thông. Thông qua việc sử dụng các kỹ thuật học sâu giống như Generative Adversarial Networks, kẻ tấn công có thể tạo video giả một cách trực quan, thực tế về mục tiêu bằng cách hoán đổi khuôn mặt trong video này với khuôn mặt khác. Tương ứng với đó, một số phương pháp đã được phát triển để phát hiện và chống lại các video deepfake này. Deepfakes là một công nghệ rất mạnh mẽ và nguy hiểm, tuy nhiên, việc sử dụng chúng vẫn còn hạn chế. Tạo video giả thường yêu cầu kẻ tấn công có kỹ năng và hầu hết các thuật toán

deepfake cũng yêu cầu một lượng lớn dữ liệu, bao gồm cả hình ảnh và video của mục tiêu.

Trong khi nhiều nghiên cứu được nhắm mục tiêu vào những kỹ thuật tiên tiến, các kỹ thuật cũ, đơn giản hơn lại không được kiểm tra, không có phương tiện phát hiện. Các thao tác chỉnh sửa video như cắt xén, nối và điều chỉnh tốc độ vẫn có thể dẫn đến các cuộc tấn công hiệu quả. Những cuộc tấn công có thể được thực hiện bởi hầu hết các phần mềm chỉnh sửa video.

Trong đề tài này, học viên sẽ nghiên cứu đánh giá một số cách tiếp cận phát hiện video bị chỉnh sửa như sau: phát hiện dựa trên đặc trưng điểm ảnh mức thấp, phát hiện dựa trên đặc trưng luồng video và phát hiện dựa trên đặc trưng audio của luồng đa phương tiện.

Đề tài sẽ đánh giá, so sánh hiệu năng và độ chính xác của từng cách tiếp cận làm cơ sở cho việc khuyến nghị sử dụng các kỹ thuật khác nhau cho từng trường hợp sử dụng cụ thể.

3. Mục đích nghiên cứu

- Rèn luyện phương pháp và khả năng nghiên cứu.
- Nghiên cứu đặc trưng video cắt ghép.
- Nghiên cứu một số thuật toán phân tích và xử lý hình ảnh.
- Ứng dụng trong một bài toán cụ thể.

4. Đối tượng và phạm vi nghiên cứu

Đối tượng và phạm vi nghiên cứu của luận văn bao gồm:

- Bài toán phát hiện điểm cắt, ghép trong video.
- Các thuật toán, phương pháp phân tích và xử lý hình ảnh.

5. Phương pháp nghiên cứu

- Phương pháp lý thuyết: Khảo sát, phân tích các tài liệu khoa học liên quan đến các thuật toán và bài toán phát hiện điểm cắt, ghép trong video.

- Phương pháp thực nghiệm: Sử dụng các công cụ, phần mềm để thử nghiệm và đánh giá hiệu quả của các thuật toán đề xuất.

Chương 1 - TỔNG QUAN VỀ BÀI TOÁN PHÁT HIỆN ĐIỂM CẮT, GHÉP TRONG VIDEO

1.1. Đặt vấn đề bài toán

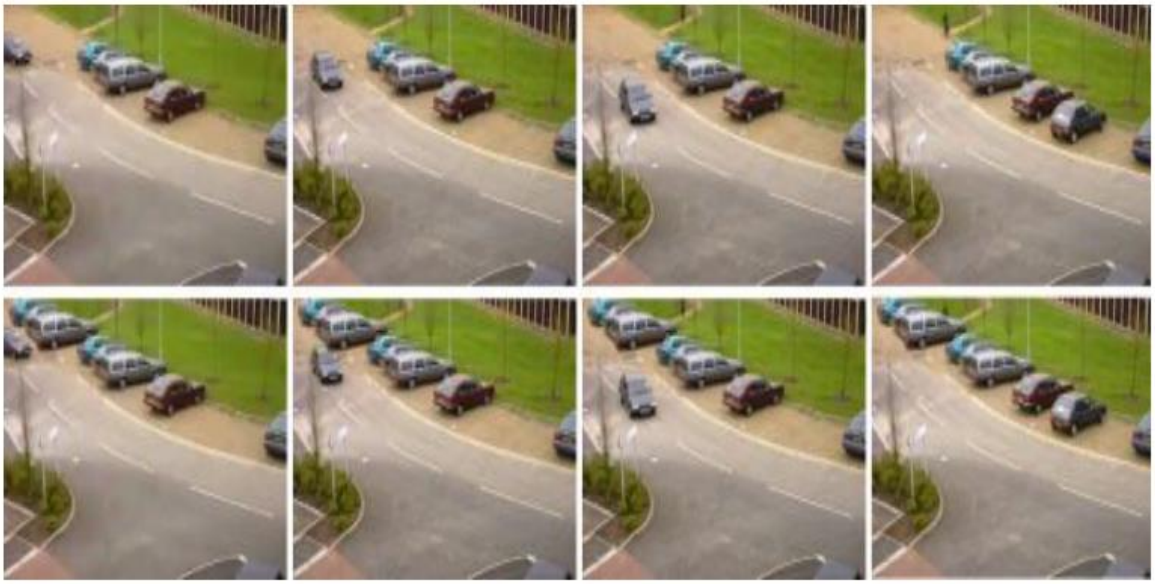
Ngày nay, sự phát triển nhanh chóng của mạng Internet kèm theo khối lượng dữ liệu khổng lồ, đa dạng và tăng trưởng không ngừng. Với sự xuất hiện, và phát triển của công nghệ mạng, người sử dụng ngày càng tăng lên, dữ liệu trên mạng internet đã trở thành một cơ sở dữ liệu phi cấu trúc lớn nhất mà con người có thể truy cập. Bắt đầu từ năm 1990, World Wide Web đã phát triển về quy mô theo cấp số nhân. Tính đến tháng 01/2021, thế giới có khoảng 4,66 tỷ người dùng Internet, chiếm 59,5% dân số thế giới [18]. Ước tính cứ mỗi ngày có hơn 2 Exabyte (10^{18} byte) dữ liệu được tạo ra trên Internet, mỗi phút có 4,2 triệu câu lệnh tìm kiếm Google; trên Facebook, có thêm 400 người dùng mới, hơn 200.000 bức ảnh được tải lên; trên Twitter, có 480.000 tài khoản được xây dựng; trên Youtube, **72 tiếng** video được tải lên, **4,7 triệu video** được xem [23]… Năm 2019, ước tính WWW chứa khoảng 4,4 Zettabytes (1 ZB = 10007 byte = 1021 byte = 10000000000000000000 byte = 1000 Exabyte = 1 Triệu Petabyte = 11000000000 (số) Terabyte = 11,000,000,000,000 Gigabyte) tài liệu web có thể lập chỉ mục công khai trải khắp thế giới trên hàng nghìn máy chủ, ước tính đến năm 2025 con số đó lên tới **175** ZB.

Đối với dữ liệu trên mạng internet, chủ yếu là nội dung do người dùng tạo (UGC), trong đó, dữ liệu là video được quay bằng thiết bị cầm tay, thiết bị điều khiển từ xa, như: điện thoại thông minh, camera, flycam… của người dùng ngày càng chiếm khối lượng lớn. Mọi người có thể chỉnh sửa video cho nhiều mục đích khác nhau, kể cả ủng hộ vấn đề chính trị hoặc giải trí, nhưng những video giả mạo như vậy đặt ra một thách thức lớn cho các tổ chức tin tức, vì việc đăng tải các video giả mạo có thể gây tổn hại nghiêm trọng đến danh tiếng, quyền lợi, sức khỏe của các tổ chức, cá nhân và thậm chí là cả xã hội. Điều này tạo ra nhu cầu cấp thiết về các công cụ có thể hỗ trợ các chuyên gia xác định và tránh nội dung video bị giả mạo. Các video này có thể chứa nội dung thực được quay tại hiện trường liên quan

tới các sự kiện và thường không chứa việc chuyển cảnh quay nào như trong các video, clip, film chuyên nghiệp, mà chỉ bao gồm một cảnh quay duy nhất. Đây là một khía cạnh quan trọng, vì một video chứa nhiều cảnh là video đã được chỉnh sửa, điều này có thể làm giảm giá trị của video đó khi được xem xét để trở thành một tài liệu chứng cứ phục vụ điều tra. Các video thường được tải lên trên các nền tảng chia sẻ phương tiện truyền thông xã hội (ví dụ: Facebook, YouTube), có nghĩa là chúng thường ở định dạng H.264 và có độ phân giải thấp và được lượng tử hóa tương đối mạnh.

Tình hình trên đã đặt ra nhiều yêu cầu đối với việc phân tích, xử lý video phục vụ công tác điều tra, giám định chứng cứ, một trong những yêu cầu nổi bật là cung cấp các công nghệ hiện đại nhất để hỗ trợ phân tích giám định video, đặc biệt là phát hiện và xử lý cục bộ các thao tác chỉnh sửa đối với video. Yêu cầu này không chỉ ngày càng cấp thiết tại Việt Nam mà còn là yêu cầu chung của nhiều nước trên thế giới. Khi nhắc đến việc phát hiện các hoạt động chỉnh sửa đối với video đồng nghĩa với việc chúng ta đề cập đến nhiệm vụ sử dụng các thuật toán phân tích video để phát hiện xem video có bị giả mạo bởi các phần mềm xử lý video hay không và nếu có, cần đưa ra các thông tin cụ thể về quá trình giả mạo (ví dụ: vị trí trong video mà giả mạo nằm ở đâu và loại giả mạo đã diễn ra).

Việc phát hiện thao tác độc hại trong các phương tiện kỹ thuật số vẫn còn nhiều hạn chế, việc phân biệt dấu vết chỉnh sửa, cắt ghép so với hình ảnh gốc ngày càng trở nên khó khăn khi các phương pháp giả mạo hình ảnh tinh vi mới được xuất hiện và phổ biến. Vì các công cụ giả mạo ngày càng thông minh, nên một hệ thống phát hiện giả mạo kỹ thuật số đáng tin cậy đang ngày càng trở nên quan trọng trong các lĩnh vực an ninh công cộng, cũng như đối với các lĩnh vực khác, như: điều tra tội phạm, pháp y, dịch vụ tình báo, bảo hiểm, báo chí, nghiên cứu khoa học, hình ảnh y tế và giám sát... Hình 1.1 là một ví dụ cụ thể, cho thấy việc đối tượng đã sao chép một số ô tô và dán vào các khung giống nhau, nhằm che đi một số thông tin gốc; trong đó, hàng trên hiển thị khung video gốc và hàng dưới hiển thị phiên bản bị giả mạo tương ứng.



Hình 1.1. Ví dụ trùng lặp đối tượng (frame gốc: trái; frame giả mạo: phải)

Tuy nhiên, các hành vi chỉnh sửa hình ảnh không phải lúc nào cũng độc hại đối với việc giám định video [13]. Bên cạnh những trường hợp có thể xảy ra như chèn hoặc xóa người, đồ vật quan trọng, có thể làm thay đổi nội dung của video và đây là những trường hợp mà giám định video đề tài chủ yếu nhắm đến, còn có rất nhiều kiểu giả mạo khác có thể diễn ra trên video nhưng không ảnh hưởng lớn tới tính chính xác của chứng cứ. Chúng có thể bao gồm các hoạt động như như điều chỉnh độ sắc nét hoặc màu sắc vì lý do thẩm mỹ cho toàn bộ video hoặc việc bổ sung các biểu tượng và hình mờ trên video. Tất nhiên, các bước xử lý hậu kỳ như vậy theo ngữ cảnh thực tế làm giảm phần nào tính chính xác và hiệu quả của video, nhưng trong những trường hợp như vậy, video vẫn là bằng chứng khả dụng duy nhất về hành vi vi phạm, chúng vẫn luôn là tài liệu vô cùng quan trọng đối với các cơ quan điều tra.

Việc phát hiện các thao tác chỉnh sửa trong video là một nhiệm vụ đầy thách thức vì các thao tác giả mạo để lại dấu vết trên video - thường không thể nhìn thấy bằng mắt thường và liên quan đến một số thuộc tính của nhiễu ảnh cơ bản hoặc các mẫu nén của video và dấu vết đó chỉ có thể được phát hiện bằng các thuật toán thích hợp nhưng hiện nay vẫn còn tồn tại nhiều phức tạp trong cách tiếp cận này. Nhìn

chung, có nhiều kiểu hành vi chỉnh sửa khác nhau có thể diễn ra, như: xóa đối tượng, sao chép đối tượng từ cùng một cảnh hoặc từ một video khác, chèn nội dung tổng hợp, chèn hoặc xóa khung, chọn khung hoặc thay đổi màu sắc/độ sáng toàn cục… mỗi loại có khả năng để lại các loại dấu vết khác nhau trên video. Hơn nữa, một vấn đề khác của bài toán thực tế là việc nén video bao gồm một số quy trình khác nhau, tất cả đều có thể phá vỡ các dấu vết giả mạo. Đặc biệt là trong trường hợp nội dung của người dùng mạng trực tuyến, chúng thường được đăng tải trên mạng xã hội, có nghĩa là chúng đã được mã hóa lại nhiều lần và thường có chất lượng thấp, do ảnh hưởng độ phân giải của camera hoặc do nhiều bước nén khi đăng. Vì vậy, để thành công, các chiến lược phát hiện chỉnh sửa video thường cần phải có khả năng phát hiện các dấu vết chỉnh sửa, cắt ghép rất yếu và rời rạc. Cuối cùng, một vấn đề làm phức tạp thêm nhiệm vụ là việc chỉnh sửa không độc hại. Như đã đề cập ở trên, đôi khi video được tạo ra có chứa các biểu tượng hoặc hình mờ do mục đích cá nhân của người quay/tạo video. Mặc dù những điều này không cấu thành việc phá hủy hoặc giả mạo video, nhưng chúng là kết quả của quá trình chỉnh sửa tương tự với quá trình giả mạo và do đó có thể dẫn đến các kết luận, đánh giá không chính xác của thuật toán hệ thống hoặc cũng có thể là một trong những yếu tố che đi các dấu vết của bộ chỉnh sửa độc hại khác.

Với những thách thức này, các nhà nghiên cứu đã và đang nghiên cứu xây dựng, triển khai nhiều hệ thống theo các hướng khác nhau nhằm hướng hỗ trợ các chuyên gia trong việc xác định các video giả mạo hoặc nâng cao hiện đại hóa lĩnh vực kỹ thuật hình sự. Các nghiên cứu trong giám định hình ảnh là tiền đề hết sức cần thiết cho mở rộng nghiên cứu các thuật toán hay "bộ lọc" nhằm xử lý video và giúp người dùng cụ thể hóa các điểm mâu thuẫn đáng ngờ trong video. Những bộ lọc này hướng tới khả năng đưa ra kết quả được hiển thị cho người dùng, giúp họ xác minh video một cách trực quan. Đi kèm với đó, việc sử dụng kiến trúc mạng nơ-ron nhân tạo (deep neuron) để phát hiện những điểm không nhất quán trong video và phân loại video là "gốc" hoặc bị giả mạo vào xây dựng hệ thống tự động

hóa quá trình phát hiện cũng là một nội dung tất yếu của công nghệ tự động hóa, một bước tiến của Trí tuệ nhân tạo (AI).

1.2. Một số nội dung cơ bản liên quan bài toán

Giám định hình ảnh và video về cơ bản là các lĩnh vực phụ của xử lý hình ảnh và video, do đó một số khái niệm từ các lĩnh vực xử lý hình ảnh/video đặc biệt quan trọng đối với nhiệm vụ của đề tài.

- Một hình ảnh (hoặc khung hình - frame) có thể được coi là một mảng 2 chiều của các bộ giá trị màu (R, G, B), tuy nhiên, nội dung màu thực tế của hình ảnh thường không liên quan đến giám định. Thay vào đó, chúng ta thường quan tâm đến các đặc điểm khác ít nổi bật hơn, như độ nhiễu, màu sắc được chuẩn hóa độ chói sáng hoặc độ sắc nét của hình ảnh.

- Giới hạn nhiễu hình ảnh (*image noise*) đề cập đến sự thay đổi ngẫu nhiên của thông tin về độ sáng hoặc màu sắc, nói chung là sự kết hợp của các đặc tính vật lý của thiết bị chụp (như cấu trúc của ống kính) và độ nén hình ảnh (trong trường hợp nén bị mất là tiêu chuẩn). Một cách để loại bỏ nhiễu hình ảnh là loại bỏ phiên bản được lọc nhiễu thấp, phần còn lại của hoạt động này có xu hướng bị chi phối bởi nhiễu hình ảnh. Trong trường hợp xử lý độ sáng thay cho việc xử lý bởi thông tin màu sắc của hình ảnh, thì chúng ta gọi là đầu ra của phương pháp đó là nhiễu độ sáng (*luminance noise*) [13].

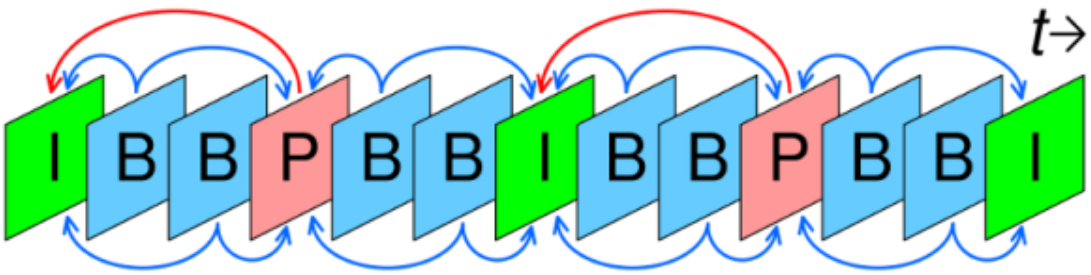
- Một vấn đề thường gặp khác của xử lý hình ảnh là sự nhạy bén (*acuity*) hoặc sắc nét (*sharpness*), chúng là sự kết hợp của độ tập trung, khả năng hiển thị và chất lượng hình ảnh; có thể được tách biệt bằng cách sử dụng bộ lọc thông cao.

- Đối với video, vấn đề nén MPEG cũng rất quan trọng đối với giám định. Nén MPEG có nhiều loại, như: MPEG-1, MPEG-2, MPEG-4 Part 2 và MPEG-4 part 10, còn được gọi là AVC hoặc H.264; về cơ bản chúng dựa trên sự khác biệt giữa các khung được mã hóa chỉ sử dụng thông tin chứa bên trong chúng (còn được gọi là nén nội khung) và các khung được mã hóa bằng cách sử dụng thông tin từ các khung khác trong video (được gọi là nén liên khung).

+ Nén nội khung về cơ bản là nén hình ảnh dựa trên các thuật toán tương tự như mã hóa JPEG.

+ Khái niệm mã hóa liên khung phức tạp hơn. Cần đưa ra các khung khác trong chuỗi, thuật toán nén thực hiện liên kết khối giữa các khung này và khung được mã hóa. Các vec-tơ liên kết các khối này được gọi là vectơ chuyển động, bên cạnh việc cung cấp cách tái tạo khung bằng cách sử dụng các phần tương tự từ các khung khác, cũng có thể cung cấp ước tính sơ bộ về các dạng chuyển động trong video, bằng cách nghiên cứu sự dịch chuyển của các đối tượng theo thời gian. Việc tái tạo khung được thực hiện bằng cách kết hợp các khối bù chuyển động từ các hệ quy chiếu, với một hình ảnh dư được thêm vào đó để tạo ra khung cuối cùng.

Các khung hình trong video được mã hóa MPEG được gắn nhãn các khung (frame) I, P hoặc B, tùy thuộc vào bảng mã của chúng. Mã hóa nội khung, mã hóa liên khung P chỉ sử dụng dữ liệu từ các khung trước đó, trong khi mã hóa liên khung hai hướng B sử dụng dữ liệu từ cả các khung trước đó và kế tiếp. Trong một video, chúng được sắp xếp theo Nhóm các hình ảnh (GOP), bắt đầu với khung I và chứa các khung P và B (Hình 1.2). Khoảng cách giữa hai I là độ dài GOP, được xác định trong các bảng mã trước đó nhưng có thể khác nhau ở các định dạng hiện đại. Tương tự, các định dạng hiện đại cho phép nhiều khả năng hơn trong các khía cạnh khác của mã hóa, chẳng hạn như kích thước và hình dạng khối, có nghĩa là các thuật toán có quy định chính xác về hoạt động của thuật toán (ví dụ: kích thước GOP cố định) sẽ không hoạt động trên các định dạng hiện đại.



Hình 1.2. Ví dụ 02 Nhóm các hình ảnh GOP

1.3. Nghiên cứu, ứng dụng hiện nay về phát hiện điểm cắt ghép trong video

Sự phát triển của công nghệ gần đây đã làm tăng lượng dữ liệu trực quan, hàng tỷ hình ảnh và video được tạo ra mỗi ngày trên web và mạng xã hội theo cấp số nhân. Các trang web truyền thông xã hội đang đóng một vai trò quan trọng hơn trong cuộc sống hàng ngày của chúng ta; Facebook, Twitter, YouTube và Instagram là những trang web trực tuyến phổ biến nhất cho phép mọi người tải lên và chia sẻ hàng trăm triệu bức ảnh. Chúng giúp người dùng thể hiện bản thân, kết bạn mới và chia sẻ sở thích cũng như ý tưởng của họ với những người khác; đồng thời, sự tác động tới đời sống xã hội và yếu tố chính trị của các phương tiện truyền thông phổ biến là không thể nghi ngờ, đặc biệt là với sự đóng góp của mạng xã hội trong việc định hình chính trị và xã hội như hiện nay trên thế giới. Để làm cho tin tức trực tuyến trở nên hấp dẫn hơn và dễ tiếp cận hơn đối với người xem, hầu hết chúng đều được gắn với nhiều hình ảnh hoặc video. Chúng cũng đại diện cho một phần đáng kể thông tin được lưu hành trong giao tiếp hàng ngày của chúng ta, ví dụ như báo chí và các trang web xã hội. Thông tin với nội dung đa phương tiện cũng được phổ biến nhanh chóng, việc đảm bảo tính toàn vẹn và tính xác thực của khối lượng dữ liệu khổng lồ trước khi sử dụng chúng trong nhiều tình huống tố tụng ngày càng quan trọng hơn [27]. Tuy nhiên, bên cạnh những lợi ích của tiến bộ công nghệ, nó cũng có thể gây ra nhiều rủi ro, đặc biệt là những rủi ro liên quan đến hệ thống xã hội và an toàn của con người. Gần đây, nhiều tin tức giả đã được thông báo rộng rãi trên phương tiện truyền thông xã hội về virus Corona (COVID-19). Thông tin về các biện pháp khắc phục sai lầm và thuyết âm mưu đã ảnh hưởng đến Internet với một loạt thông tin sai lệch, nguy hiểm. Thông qua các phương tiện truyền thông, thông tin sai sự thật có thể lan truyền nhanh hơn và dễ dàng hơn trên mạng xã hội và Internet. Do đó, sự phổ biến của những thông tin không chính xác vừa không hữu ích hoặc thậm chí có tác động tiêu cực rất lớn tới sức khỏe cộng đồng và làm trầm trọng thêm tình trạng bất ổn và chia rẽ xã hội. Ví dụ: vào tháng 01/2020, một số lượng lớn các tin đồn dưới dạng hình ảnh và video clip lan truyền trên mạng liên

quan đến virus COVID-19 khiến nhiệm vụ phân biệt giữa các thông tin, tin tức thật và giả ngày càng trở nên khó khăn. Vì vậy, Tổ chức Y tế Thế giới (WHO) đã phải đưa ra cảnh báo đối với mọi người với danh sách thông tin sai lệch về virus Corona.

Ngày nay, giám định đa phương tiện kỹ thuật số đã trở thành một lĩnh vực nghiên cứu mới nổi, nhận được sự chú ý đáng kể nhằm xác định nguồn gốc và tính xác thực của phương tiện kỹ thuật số. Tính xác thực của hình ảnh rất quan trọng trong nhiều lĩnh vực xã hội, chẳng hạn như: *trong lĩnh vực y tế,* các bác sĩ đưa ra các quyết định quan trọng dựa trên hình ảnh kỹ thuật số; *trong các cơ quan thực thi pháp luật và trong tố tụng hình sự,* tính chính xác của các bức ảnh có một vai trò thiết yếu để chúng có thể được sử dụng làm bằng chứng. Trong thời đại kỹ thuật số ngày nay, sự phát triển nhanh chóng của các công cụ chỉnh sửa mạnh mẽ và chi phí thấp tạo điều kiện thuận lợi cho việc cắt ghép video/hình ảnh trên các phương tiện kỹ thuật số, như thêm hoặc bớt các phần và đối tượng khỏi hình ảnh và video, nhờ đó có thể ít hoặc không để lại dấu vết của việc cắt ghép, chỉnh sửa. Sau đó, phương tiện bị chỉnh sửa, cắt ghép này sẽ lan truyền nhanh chóng và có thể gây ra những hậu quả nghiêm trọng, trên cả quy mô quốc gia và quốc tế. Hơn nữa, để đảm bảo tính toàn vẹn và tính xác thực của chúng là vô cùng khó khăn, như trong Hình 1.3, đại diện cho một trường hợp giả mạo thường gặp. Với những tiến bộ nhanh chóng của máy ảnh kỹ thuật số độ phân giải cao và tiện ích của phần mềm chỉnh sửa phức tạp, chẳng hạn như Adobe Photoshop, Pixar và Corel PaintShop, người dùng có thể dễ dàng sửa đổi nội dung của ảnh mà không để lại bất kỳ dấu hiệu chỉnh sửa cảm quan rõ ràng nào, chúng đang vô tình làm mờ ranh giới giữa nội dung *thật* và *giả*. Việc sử dụng không đúng các công cụ chỉnh sửa như vậy khiến các video giả mạo và xuyên tạc trên mạng xã hội đang trở thành một vấn đề ngày càng nghiêm trọng. Thật vậy, những kẻ làm giả video liên tục cố gắng khai thác các công cụ này để che giấu hình ảnh và video thực, sau đó sử dụng chúng để diễn giải sai thông tin có thể lan truyền rất nhanh và có thể gây ra hậu quả vô cùng lớn. Chúng cũng có thể dẫn đến các vấn đề phát triển nhanh chóng như làm giảm độ tin cậy trên nhiều ứng dụng

thực tế, khiến người xem rất khó đánh giá tính xác thực của một hình ảnh hoặc video nhất định.

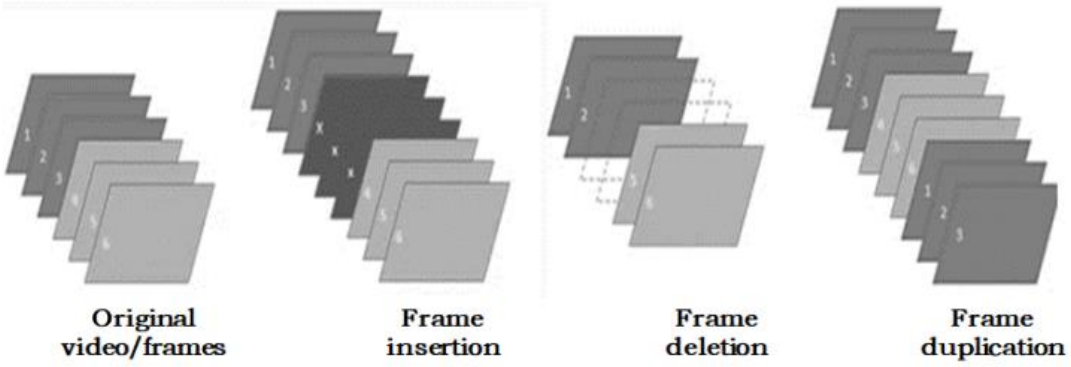
Việc chỉnh sửa phương tiện truyền thông kỹ thuật số thường được gọi là ***giả mạo kỹ thuật số*** là nỗi lo ngại lớn đối với cá nhân (như chuỗi video giả mạo của những người nổi tiếng), đối với xã hội (như hình ảnh giả mạo khiêu khích nhằm vào một số sắc tộc hoặc tôn giáo nhất định), đối với báo chí, các công ty bảo hiểm và các tạp chí khoa học... Giả mạo trở thành nỗi lo đối với các chính phủ, các doanh nghiệp công và tư nhân và đối với cuộc sống riêng tư của các cá nhân. Do đó, thế giới đang đối mặt với một thách thức nghiêm trọng cần giải quyết ngay là vấn nạn phát tán ảnh và video lừa đảo.

Hình 1.3. Ảnh gốc (trái) và ảnh giả mạo (phải)

Gần đây, một số nhà nghiên cứu khoa học đã xem xét tính xác thực của phương tiện truyền thông nhưng do khối lượng đa phương tiện khổng lồ và phức tạp cần phân tích khiến việc xây dựng thuật toán phát hiện giả mạo đa phương tiện trở nên khó khăn. Nghiên cứu trong lĩnh vực này chưa đưa ra được các giải pháp mạnh mẽ và phổ biến, đến nay vẫn cần nhiều những nghiên cứu, đóng góp sâu rộng hơn. Trong những năm gần đây, hầu hết các nỗ lực đã được dành cho việc phát hiện *giả mạo tĩnh*, việc phát hiện *giả mạo động* đã không nhận được nhiều sự chú ý vì sự phức tạp của phân tích cảnh động và chi phí tính toán, vấn đề này trở nên khó khăn

hơn với giám định video. Trên thực tế, các vấn đề nghiêm trọng đối với việc phát hiện giả mạo video, như: sự phức tạp của phân tích cảnh động, chi phí tính toán, sự hiện diện của việc chuyển cảnh, những thay đổi về phối cảnh, tỷ lệ, điều kiện ánh sáng khác nhau và việc khai thác các đối tượng theo không gian - thời gian (ví dụ: màu sắc, kết cấu, hình dạng, cấu trúc, bố cục và chuyển động). Tất cả những vấn đề này thúc đẩy nhu cầu nghiên cứu lĩnh vực nghiên cứu nóng bỏng này.

Một số công trình khoa học có liên quan đã được phát triển để phát hiện video giả mạo hoặc có khả năng phát hiện các đối tượng hoặc khung hình đáng ngờ dựa trên các đặc điểm của video kỹ thuật số. Một số phương pháp được triển khai tập trung vào việc xác định giả mạo giữa các khung hoặc nội khung [20]. Các phương pháp dựa trên xem xét nội khung có thể thực hiện trong miền không gian hoặc không gian - thời gian (như sao chép - di chuyển hoặc nối khung). Các phương pháp dựa trên liên khung (Hình 1.4) diễn ra trong miền thời gian (như chèn, loại bỏ và sao chép khung). Một trong những công trình tiên phong trong lĩnh vực này đã xử lý việc phát hiện trùng lặp khung [33], bằng cách tính đến thông tin tương quan giữa các khung liên tiếp. Các loại tấn công và giả mạo khác nhau có thể xảy ra để thay đổi và xóa bằng chứng video. Do đó, các manh mối hiệu quả cần được khai thác để phát hiện ra những sự giả mạo này, ví dụ bao gồm: tốc độ và sự không nhất quán về mặt vật lý [5]; phần dư chuyển động [35]; và các tính năng đường bao thống kê [4].



Hình 1.4. Ví dụ về việc giả mạo liên khung.

Nhìn chung, video giả mạo có thể được phát hiện bằng cách xác minh các thay đổi về không gian, chẳng hạn như nén khung hình [14] [22] hoặc các phương thức về mặt thời gian như thêm hoặc xóa khung [2] [12]. Trong số các kỹ thuật của giám định thụ động, nén kép là một trong những manh mối quan trọng để phát hiện giả mạo video. Khi xử lý video nén, những kẻ tấn công làm theo các bước nhất định để sửa đổi video này bằng cách: đầu tiên, giải mã video này; sau đó thao tác chỉnh sửa và cuối cùng giải nén nó. Rõ ràng, kịch bản này sẽ để lại dấu vết có thể được khai thác làm thông tin có giá trị để phân tích giám định. Một số nghiên cứu đã giải quyết vấn đề phát hiện nén kép như dựa trên việc sử dụng các đặc trưng không gian - thời gian được đánh giá trên cơ sở trường vectơ chuyển động cục bộ [15].

Một số nghiên cứu tập trung vào việc phát hiện sự trùng lặp khung, ví dụ, trong [34], các tác giả khai thác mối tương quan của các đặc điểm phân tách giá trị kỳ dị giữa các khung gốc và khung đáng ngờ, việc giả mạo sao chép khung được phát hiện bằng cách sử dụng phương pháp dựa trên phân tích tương tự. Ngoài ra, các đặc điểm dư chuyển động trong mỗi khung có thể được sử dụng để xác định các khung bị chỉnh sửa, giả mạo. Một kỹ thuật thụ động khác dựa trên việc trích xuất các đặc trưng thống kê và phân loại của các đặc điểm này thành các mẫu dương tính hoặc mẫu âm tính [26]. Các tính năng trực quan được bắt nguồn từ thời điểm dựa trên sóng và cường độ gradient trung bình, quá trình trích xuất dựa trên khái niệm về ranh giới đối tượng có chiều rộng có thể điều chỉnh (AWOB). Việc phát hiện trùng lặp khung hình cũng được xử lý khác nhau, đặc biệt là với bộ mô tả SIFT và mô hình bag-of-words (BoW) [32]. Kỹ thuật này chỉ có thể phát hiện việc tạo ra các khung sao chép chứ không phát hiện được các hình thức tấn công khác. Các công trình nghiên cứu khác đã giải quyết đồng thời nhiều loại tấn công khác nhau như xóa khung và chèn khung bằng cách sử dụng biểu đồ của các tính năng gradient có định hướng (HOG). Các nhà nghiên cứu khai thác cái gọi là luồng video để trích xuất rìa hình ảnh và sau đó xác định vị trí của cả thao tác nhân bản khung hình.

Việc khai thác các đặc điểm không gian-thời gian hiệu quả vẫn là thách thức chính đối với hầu hết các nhà nghiên cứu để xác định các khung hình sao chép với

độ chính xác cao [27]. Ví dụ: phép phân tách giá trị số ít (SVD) được thực hiện cùng với phép đo độ tương tự Euclid trong [34]; độ lệch chuẩn của các khung hình dư được sử dụng để chọn một số khung hình từ chuỗi video và sau đó giá trị entropy của Biến đổi Cosine rời rạc (DCT) được khai thác để phát hiện sự trùng lặp giữa các khung hình [9]. Trong [28], các tác giả đã sử dụng DCT để tạo một tập hợp các đặc trưng cho mỗi khung và sau đó để phát hiện ra sự hiện diện của giả mạo bằng cách sử dụng hệ số tương quan. Phương pháp này cho kết quả tốt nhưng thời gian tính toán tương đối lớn.

Video cũng có thể được giả mạo bằng thao tác nối thời gian. Để giải quyết loại giả mạo này, một máy dò đã được thiết kế trong, nó đánh giá một video có được nội suy theo thời gian hay không bằng cách tính toán mối tương quan thời gian giữa các khung hình video [1]. Sau đó, trình phát hiện này đã được cải tiến bằng cách tận dụng cường độ cạnh để xác định sự hiện diện của việc thay đổi tốc độ khung hình video. Các tác giả cũng đã nghiên cứu đường trung bình động thích ứng Kaufman (KAMA) để tách các khung xác thực khỏi các khung nội suy. Các manh mối và dấu vết khác, đặc biệt là mối tương quan dựa trên nhiễu video, cũng đã được kiểm tra để tiến hành phát hiện video giả mạo dựa trên việc khai thác nhiễu được trích xuất như một đặc đặc trưng mạnh mẽ và sử dụng kỹ thuật tương quan mức khối [16]. Họ mô hình hóa sự phân bố tương quan của dư lượng nhiễu theo thời gian trong một video giả mạo dưới dạng mô hình hỗn hợp Gaussian (GMM). Tuy nhiên, cách tiếp cận của họ phụ thuộc rất nhiều vào kỹ thuật khử nhiễu. Khi cường độ nhiễu của vùng gốc và vùng bị xáo trộn khác nhau, nó không thể giảm nhiễu một cách chính xác và có thể bỏ sót một số giả mạo do sai số tính toán dư nhiễu. Các mô hình hỗn hợp dựa trên Gaussian (GMM) thông thường là các công cụ phổ biến cho các kết quả chấp nhận được để lập mô hình dữ liệu đơn biến; tuy nhiên, chúng không có nhiều hình dạng phức tạp khác nhau. Phương pháp thứ hai có thể cung cấp nhiều khả năng hơn để thích ứng tốt hơn với dạng dữ liệu không phải Gaussian là phân phối Gaussian thông thường (GMM). Một cách tiếp cận khác, trong đó chức năng mức nhiễu (NLF) được sử dụng để phát hiện các vùng khả nghi trong cảnh

tĩnh được ghi lại từ video. Các tác giả xử lý NLF tuyến tính và phi tuyến như sự không nhất quán của nhiễu để phát hiện các vùng giả mạo [19].

Gần đây, một số kỹ thuật phát hiện giả mạo video tự động đã được triển khai, trong đó, có những cách tiếp cận tận dụng các mô hình thống kê được áp dụng thành công. Việc chuyển đổi tốc độ khung hình bù theo chuyển động cũng được khai thác cho các mục đích phát hiện giả mạo như làm giả tốc độ khung hình. Vấn đề này cũng được xử lý, trong đó tín hiệu dư được coi là dấu hiệu để xác định vị trí các khung giả mạo nội suy [6]. Thời điểm trên xung dao động wavelet và cường độ gradient trung bình cũng được ước tính cùng với khái niệm về ranh giới đối tượng có chiều rộng có thể điều chỉnh (AWOB) và phân loại SVM để xác định các mẫu dương tính (video gốc) và mẫu âm tính (video giả mạo) [26].

Có thể thấy, các nghiên cứu hiện nay trong lĩnh vực giám định video đã đạt được nhiều thành tựu lớn, kết quả khả quan. Tuy nhiên, còn tồn tại một số khó khăn như: hiệu quả khử nhiễu thấp, chưa hoạt động hiệu quả trên video chất lượng cao, khó để định vị tất cả các khung hình nội suy và không thể khôi phục video đã bị chỉnh sửa, cắt ghép trong nhiều trường hợp.

Chương 2 - THUẬT TOÁN VÀ MÔ HÌNH HỆ THỐNG TỰ ĐỘNG PHÁT HIỆN ĐIỂM CẮT, GHÉP TRONG VIDEO

2.1. Các đặc trưng của video bị cắt ghép, giả mạo

Video là một tập hợp của các chuỗi khung hình/hình ảnh kết hợp với các kỹ thuật nén khác nhau, do đó, ở một mức độ nào đó các loại giả mạo video có thể có những thông tin sai lệch tương tự như các loại giả mạo trong hình ảnh, như: có thể gặp phải các thao tác sao chép chuyển động, ghép nối, nội khung hoặc chỉnh sửa toàn bộ video như thay đổi độ sáng hoặc độ nét. Tuy nhiên, một điểm khác biệt quan trọng trong giám định video là các thao tác giả mạo có thể tác động đến phương diện thời gian của video, ví dụ như chỉnh sửa ghép nối thường là việc chèn video khác bao gồm nhiều khung chứa hình ảnh mô tả vật thể mới đang chuyển động vào video gốc; tương tự, quá trình copy-move có thể bị dịch chuyển mặt thời gian, tức là một đối tượng của video từ một số khung hình xuất hiện lại trong các khung hình khác hoặc bị dịch chuyển theo không gian, tức là một đối tượng từ một khung hình xuất hiện lại ở nơi khác trên cùng một khung hình. Hơn nữa, tồn tại một loại giả mạo chỉ có thể có trong video, cụ thể là giả mạo giữa các khung hình, bao gồm chèn hoặc xóa khung.

Ngoài ra, các thuật toán giám định hình ảnh dựa trên định dạng ảnh JPEG là không đủ để phát hiện hoặc xác định vị trí các điểm giả mạo trong video. Lý do chính cho điều này là một video không chỉ là một chuỗi hình ảnh; việc nén MPEG - đây là định dạng video phổ biến nhất hiện nay - mã hóa thông tin bằng cách khai thác mối tương quan thời gian giữa các khung hình, về cơ bản là tái tạo lại hầu hết các khung hình bằng cách kết hợp các khối từ các khung hình khác với một hình ảnh dư. Quá trình này về cơ bản phá hủy các dấu vết mà các thuật toán dựa trên hình ảnh nhằm mục đích phát hiện. Hơn nữa, việc yêu cầu và giải nén được thực hiện bởi các nền tảng trực tuyến như YouTube, Facebook và Twitter gây khó khăn hơn nhiều đối với giám định các dấu vết giả mạo nhỏ, khó phát hiện so với các thuật toán giải nén tương ứng cho hình ảnh. Do đó, việc phát hiện giả mạo video đòi hỏi sự phát triển của các thuật toán cụ thể hướng mục tiêu đến đối tượng là các video.

Hơn nữa, các thuật toán được thiết kế cho MPEG-2 thường sẽ bị lỗi khi gặp phải các video MPEG-4/H.264, đây là định dạng phổ biến cho các video trực tuyến hiện nay. Vì vậy, khi khảo sát tình trạng kỹ thuật, có thể sử dụng một phương pháp phân loại tương tự để kiểm tra hình ảnh cho các thuật toán dựa trên video. Có thể tìm thấy một số lượng lớn các phương pháp giám định tích cực, tuy nhiên, các phương pháp này không áp dụng được trong khá nhiều trường hợp, nơi chúng ta không kiểm soát được quá trình quay video. Như đã đề cập ở trên, giám định video tự động có thể được tổ chức theo cấu trúc tương tự như giám định hình ảnh tự động, liên quan đến loại giả mạo nhằm phát hiện: ghép nối/chèn đối tượng, di chuyển bản sao/nhân bản, chỉnh sửa toàn bộ video và chèn/xóa khung hình.

Do đó, các phương pháp tiếp cận giám định video được đề xuất có thể được phân theo ba loại: phát hiện lượng tử hóa kép/nhiều, phát hiện giả mạo giữa các khung và phát hiện giả mạo vùng.

- Trong trường hợp đầu tiên, các hệ thống cố gắng phát hiện xem một video hoặc các phần của nó đã được lượng tử hóa nhiều lần hay chưa [30]. Một video là Nội dung do người dùng tạo (UGC) trên máy ảnh nhưng thể hiện dấu vết của nhiều phép lượng hóa thì video đó có thể đáng ngờ. Tuy nhiên, đối với UGC đáng tin cậy, các cách tiếp cận như vậy không đặc biệt phù hợp vì trong phần lớn các trường hợp, video được lấy từ các nguồn truyền thông xã hội. Do đó, cả video bị giả mạo và chưa được kiểm tra thường trải qua nhiều lần lượng tử hóa và rất khó để xác thực nếu không có quyền truy cập vào bản gốc của máy ảnh.

- Trong loại thứ hai, để phát hiện giả mạo giữa các khung, các thuật toán nhằm mục đích phát hiện các trường hợp khung mới đã được chèn thêm vào video [37]. Giả mạo giữa các khung hình là một loại giả mạo video đặc biệt, bởi vì nó có thể nhận dạng trực quan trong hầu hết các trường hợp, như: một sự thay đổi cảnh quay hoặc cắt đột ngột trong video. Có hai loại video mà sự giả mạo như vậy có thể thực sự thành công để đánh lừa người xem: ***Một là,*** trường hợp video đã có các đoạn cắt, tức là cảnh đã chỉnh sửa. Ở đó, một cảnh quay có thể bị xóa hoặc thêm

vào trong số các ảnh hiện có, nếu bản âm thanh có thể được chỉnh sửa tương ứng. ***Hai là,*** trường hợp của video CCTV hoặc các cảnh video được quay từ một camera tĩnh, ở đó, các khung hình có thể được chèn, xóa hoặc thay thế mà không gây chú ý về mặt trực quan. Tuy nhiên, ngày nay phần lớn video thường được chụp bởi các thiết bị chụp cầm tay, gồm các ảnh đơn chưa bị chỉnh sửa, việc chèn giữa các khung không thể được áp dụng mà không gây chú ý. Vì vậy, khi giám định video, chúng ta có thể khái niệm đây như một phần mở rộng của giám định hình ảnh, có thể được giải quyết bằng các giải pháp tương tự. Ví dụ: ghép nối video có thể được phát hiện dựa trên giả định rằng phần được chèn có lịch sử ghi và nén khác với video nhận nó. Tuy nhiên, các nghiên cứu thử nghiệm sơ bộ cho thấy rằng các thuật toán được thiết kế cho hình ảnh không hoạt động tốt trên video và điều này thậm chí còn áp dụng cho các thuật toán dựa trên nhiễu chung nhất [27].

- Cuối cùng, loại thứ ba phát hiện giả mạo vùng liên quan đến các trường hợp các phần của chuỗi video (ví dụ: một đối tượng) đã được chèn vào các khung của một video khác. Đây là kịch bản thường gặp nhất cho UGC. Các thuật toán phát hiện giả mạo vùng video chia sẻ nhiều nguyên tắc chung với các thuật toán phát hiện ghép nối hình ảnh. Trong cả hai trường hợp, giả định là tồn tại một số mẫu không thể nhìn thấy bằng mắt thường, do quá trình chụp hoặc nén, không hoạt động, có thể phát hiện được và có thể bị xáo trộn khi nội dung ngoài được chèn vào. Một số cách tiếp cận dựa trên thông tin về mặt không gian được trích xuất riêng từ các khung. Trong số đó, những phương pháp nổi bật nhất là sử dụng gradient có định hướng hoặc biểu đồ hệ số biến đổi Cosine rời rạc (DCT). Chúng hoạt động tốt trên những video chất lượng cao, nhưng có xu hướng không thành công ở độ nén cao hơn vì các dấu vết chỉnh sửa hầu như đã bị xóa.

Các chiến lược phát hiện giả mạo vùng khác dựa trên thành phần chuyển động của mã hóa video, lập mô hình thống kê vectơ chuyển động hoặc thống kê lỗi bù chuyển động. Các phương pháp này hoạt động tốt hơn với nền tĩnh và các đối tượng chuyển động chậm, sử dụng chuyển động để xác định hình dạng/đối tượng

cần quan tâm trong video. Tuy nhiên, những điều kiện này thường không đáp ứng được UGC.

2.2. Một số thuật toán phát hiện điểm cắt, ghép trong video và đề xuất

2.2.1. Một số thuật toán phát hiện điểm cắt, ghép trong video

2.2.1.1. Phương pháp tiếp cận dựa trên đặc trưng ảnh

Giám định hình ảnh là một lĩnh vực lâu đời hơn giám định video; với khối lượng lớn các thuật toán đã được xây dựng dựa trên khai thác các đặc trưng ảnh kỹ thuật số cùng lượng lớn các bộ dữ liệu thử nghiệm, giám định hình ảnh đang dần đạt đến độ chín khi các thuật toán hoặc các tổ hợp thuật toán đang đạt đến độ chính xác tối đa cho ứng dụng trong thế giới thực. Việc phát hiện giả mạo hình ảnh thường dựa trên việc phát hiện sự không nhất quán cục bộ trong thông tin nén JPEG, hoặc phát hiện sự không nhất quán cục bộ trong các mẫu nhiễu tần số cao do thiết bị chụp để lại (đặc biệt trong những trường hợp hình ảnh chất lượng cao, độ nén thấp). Sự tiến bộ trong giám định hình ảnh có thể đưa ra kết luận rằng các phương pháp tương tự có thể hoạt động để phát hiện video giả mạo. Cụ thể:

- Mặc dù, giám định đa phương tiện là một lĩnh vực có lịch sử nghiên cứu lâu đời và đã đạt được nhiều tiến bộ trong những thập kỷ qua, tuy nhiên, hầu hết những nghiên cứu này đều liên quan đến phân tích hình ảnh, có rất ít nghiên cứu chuyên sâu về phân tích video. Các phương pháp giám định hình ảnh thường được tổ chức theo một trong hai loại sau: *(1) Giám định tích cực,* trong đó hình mờ hoặc phần thông tin tương tự (thường không nhìn thấy) được nhúng vào hình ảnh tại thời điểm chụp, trong đó tính toàn vẹn được đảm bảo rằng hình ảnh không bị chỉnh sửa kể từ khi chụp [13] [24] [25]; và *(2) Giám định thụ động*, khi không tồn tại thông tin trước đó và việc phân tích xem một hình ảnh có bị giả mạo hay không hoàn toàn phụ thuộc vào chính nội dung hình ảnh đó. Mặc dù giám sát thu động là một nhiệm vụ khó khăn hơn nhiều, nhưng nó có liên quan nhất trong phần lớn các trường hợp sử dụng thực tế, khi chúng ta thường không có quyền truy cập vào quá trình chụp ảnh.

Một điểm khác biệt quan trọng trong các thuật toán giám định hình ảnh là phân biệt giữa phát hiện sự giả mạo và xác định vị trí điểm giả mạo [13]. Trong trường hợp đầu, liên quan phát hiện sự giả mạo, các thuật toán chỉ trả về kết quả đánh giá việc liệu hình ảnh có bị giả mạo hay không và thường trả về một con số ước tính khả năng giả mạo. Trong trường hợp thứ hai, thuật toán cố gắng thông báo cho người dùng vị trí quá trình giả mạo đã diễn ra và trả về một bản đồ tương ứng với hình dạng của hình ảnh và làm nổi bật các vùng của hình ảnh có khả năng đã bị giả mạo - ước tính xác suất trên mỗi khối hoặc trên mỗi pixel. Các phương pháp tiếp cận giám định hình ảnh thụ động có thể được phân loại theo phương thức mà chúng dự định phát hiện và xác định vị trí giả mạo. Ba nhóm chính của sự chỉnh sửa là *sao chép di chuyển (copy-move), ghép nối hoặc giả mạo nội khung và thay đổi toàn bộ hình ảnh*. Trong trường hợp đầu tiên, một phần của hình ảnh được sao chép và đặt ở vị trí khác trong ảnh, ví dụ: nền được sao chép để xóa một đối tượng, hoặc sao chép người hay một đám đông để có giả mạo về số lượng. Các thuật toán phát hiện copy-move cố gắng nắm bắt sự giả mạo bằng cách tìm kiếm các điểm tự tương đồng trong hình ảnh [29] [34]. Trong trường hợp ghép nối, một phần của hình ảnh này được đặt trong hình ảnh khác. Các thuật toán phát hiện và xác định vị trí ghép dựa trên tiền đề rằng, ở một mức độ nào đó (có thể không nhìn thấy) khu vực được ghép sẽ khác với phần còn lại của hình ảnh do lịch sử chụp và nén khác nhau của chúng. Trường hợp nội khung (in-painting), tức là khi một phần của hình ảnh bị xóa và sau đó được tự động vẽ bằng thuật toán in-painting về nguyên tắc tương tự, vì phần do máy tính tạo ra sẽ mang một đặc điểm khác với phần còn lại của hình ảnh.

Các thuật toán phát hiện những giả mạo trên có thể khai thác sự mâu thuẫn trong lịch sử nén JPEG cục bộ [10], trong các mẫu nhiễu cục bộ [21] hoặc trong các dấu vết do Mảng lọc màu (CFA) của thiết bị chụp [7] [11]. Trong nhiều trường hợp, các thuật toán này cũng có thể phát hiện các hành vi giả mạo copy-move, vì chúng cũng thường gây ra các gián đoạn cục bộ có thể phát hiện được. Đối với những trường hợp không cần thiết xác định vị trí giả mạo, các thuật toán phát hiện giả mạo kết hợp bộ lọc và học máy đã được đề xuất, đạt độ chính xác rất cao trong một số bộ

dữ liệu. Cuối cùng, các hoạt động giả mạo toàn bộ hình ảnh như thay đổi tỷ lệ, giải nén lại hoặc bộ lọc không thể xác định vị trí giả mạo, do đó thường được giải quyết bằng các thuật toán phát hiện giả mạo trên toàn ảnh [36].

Nhận xét: Mặc dù, gần đây, với sự ra đời của học sâu (deep learning), các phương pháp tiếp cận mới bắt đầu xuất hiện, cố gắng tận dụng sức mạnh của mạng nơ-ron phức hợp để xác định và phát hiện vị trí giả mạo; một cách tiếp cận là áp dụng một bước lọc trên hình ảnh, sau đó sử dụng kết hợp Mạng nơ-ron nhân tạo để phân tích đầu ra [3]; các phương pháp khác đã cố gắng kết hợp bước lọc vào mạng, thông qua việc giới thiệu Lớp kết hợp có ràng buộc, trong đó các tham số là giá trị chuẩn hóa ở mỗi lần lặp lại của quá trình học máy hứa hẹn nhiều khả năng ứng dụng giám định hình ảnh trong giám định video. Nhưng cho đến nay, phương pháp giám định video tiếp cận dựa trên đặc trưng ảnh mới chỉ có thể đúng và hoạt động hiệu quả cao nếu video chỉ đơn giản là chuỗi các khung hình, do việc nén video hiện đại là một quá trình phức tạp hơn nhiều, nó thường loại bỏ tất cả các dấu vết như lỗi máy ảnh và dấu vết nén khung hình đơn. Vì vậy, phương pháp tiếp cận trên là chưa khả thi trong ứng dụng thực tế và không đáp ứng được sự phát triển của thế giới công nghệ video hiện nay.

2.2.1.2. Phương pháp tiếp cận dựa trên đặc trưng luồng đa phương tiện

2.2.1.2.1. Sử dụng các bộ lọc số học

Thuật ngữ các bộ lọc số học đề cập đến các cách tiếp cận đại số cho phép chiếu thông tin vào một không gian đặc trưng nhằm giúp việc phân tích trong công tác giám định video dễ dàng hơn [13].

• Các bộ lọc **Q4** được sử dụng để phân tích sự phân rã của hình ảnh thông qua Biến đổi Cosine rời rạc (DCT). DCT 2 chiều chuyển đổi một ma trận N ×N khối hình ảnh thành một ma trận mới N×N khối, trong đó hệ số được tính toán dựa trên tần số của chúng. Cụ thể trong mỗi khối, hệ số đầu tiên nằm ở vị trí (0,0) đại diện cho thông tin tần số thấp nhất và giá trị của nó có liên quan đến giá trị trung

bình của toàn bộ khối, hệ số (0,1) bên cạnh nó đặc trưng cho sự thay đổi chậm từ tối sang sáng theo hướng ngang, v.v.

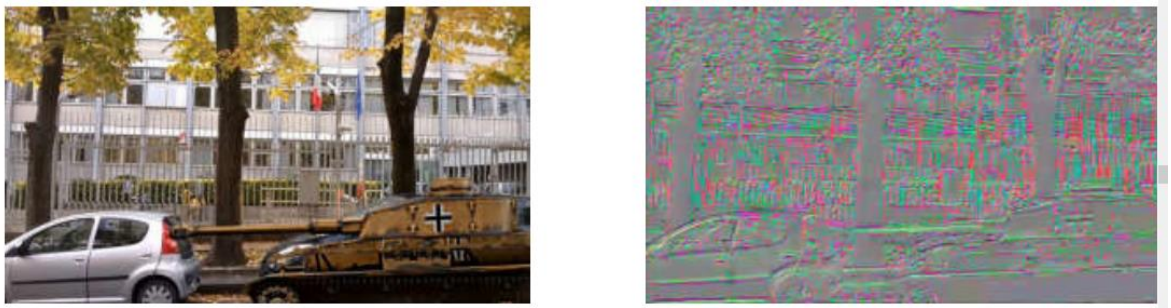
Nếu chúng ta biến đổi tất cả N×N các khối của một hình ảnh với DCT, chúng ta có thể xây dựng ví dụ như một kênh hình ảnh đơn của hệ số (0,0) ứng với mỗi khối. Hình ảnh này sau đó sẽ nhỏ hơn N lần trên mỗi thứ nguyên. Nói một cách tổng quát hơn, có thể xây dựng một hình ảnh bằng cách sử dụng các hệ số tương ứng với vị trí (i, j) của mỗi khối cho bất kỳ cặp i và j. Ngoài ra, người ta có thể tạo ra hình ảnh màu sai bằng cách chọn ba vị trí khối và sử dụng ba mảng kết quả làm kênh màu đỏ, xanh lục và xanh lam của hình ảnh đầu ra, như được minh họa bằng biểu thức hình 2.1 dưới đây:

$$\begin{pmatrix} \text{red} \\ \text{green} \\ \text{blue} \end{pmatrix} = \begin{pmatrix} \text{coefficients \#1} \\ \text{coefficients \#2} \\ \text{coefficients \#3} \end{pmatrix}.$$

Hình 2.1. Bộ chuyển đổi hệ màu của bộ lọc Q4

Để triển khai các bộ lọc Q4, có thể sử dụng các khối có kích thước 2×2 điểm ảnh. Vì hệ số tương ứng với vị trí khối (0,0) không liên quan đến xác minh và chỉ trả về một phiên bản tần số thấp của hình ảnh. Có ba hệ số có thể tạo ra một hình ảnh màu sai. Do đó, trong trường hợp này, kênh màu đỏ tương ứng với các tần số ngang (0,1), kênh màu xanh lá tương ứng với các tần số dọc (1,0) và màu xanh lam tương ứng với các tần số dọc theo hướng chéo (1,1) [13].

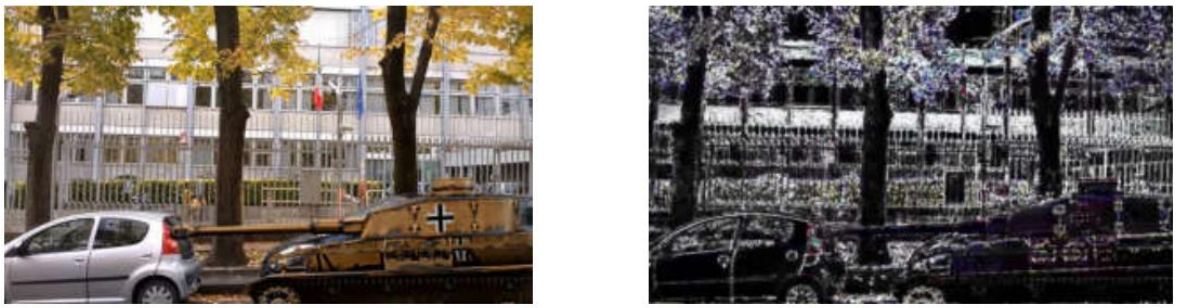
• Các bộ lọc **Chrome** chuyên dùng để phân tích nhiễu độ chói của hình ảnh. Nó làm nổi bật tính đồng nhất của nhiễu được mong đợi trong một hệ thống quan sát bình thường và được chiếu sáng tự nhiên. Nó chủ yếu dựa trên bộ lọc không tuyến tính để thu được nhiễu xung động. Do đó, các bộ lọc Chrome chủ yếu dựa trên hoạt động sau được áp dụng trên mỗi khung hình của video:



(a) (b)

Hình 2.2. Đầu ra của bộ lọc Q4 trên video xe tăng đã chỉnh sửa (a - khung đã bị chỉnh sửa, b - đầu ra bộ lọc).

Theo Hình 2.12, hình ảnh đầu ra (b) hiển thị màu đỏ tăng cường chuyển đổi theo chiều dọc (tương ứng với chuyển tiếp dọc theo các đường), màu xanh lá cây là chuyển đổi ngang và màu xanh lam là chuyển đổi theo đường chéo (chủ yếu có thể được nhìn thấy trong lá cây).



(a) (b)

Hình 2.3. Đầu ra của bộ lọc Chrome trên video xe tăng đã chỉnh sửa (a - khung đã bị chỉnh sửa, b - đầu ra của bộ lọc).

Hình ảnh (b) là ảnh màu đen và trắng nhưng vẫn còn thông tin về màu sắc do xuất phát từ bộ chuyển đổi màu liên quan nhiễu (2.1), nó cho thấy rằng nhiễu có cùng mức độ độc lập với các dải màu đầu vào.

$$I_{Chrome}(x) = |I(x) - median(W(I(x)))| \quad (2.1)$$

Trong đó I(x) biểu thị một pixel hình ảnh và W(I(x)) đại diện cho một khối 3×3 cửa sổ xung quanh pixel đó.

Bộ lọc này tương tự như thuật toán nhiễu trung bình (Median Noise) cho giám định hình ảnh, được triển khai trong Bộ công cụ giám định hình ảnh (MKLAb-ITI), trong đó bộ lọc trung bình hình ảnh được sử dụng để phát hiện sự không nhất quán trong hình ảnh. Về cơ bản, vì *nó có khả năng cô lập nhiễu tần số cao,* nên cách tiếp cận này cung cấp cái nhìn tổng quan về toàn bộ khung hình, các vị trí có dấu vết nhiễu khác nhau có thể được phát hiện và xác định là nổi bật hơn so với phần còn lại của khung hình.

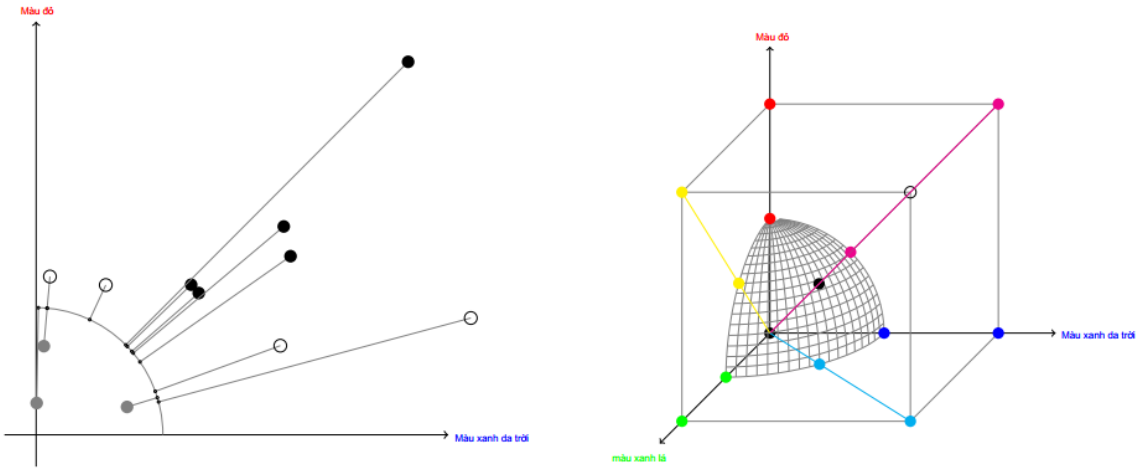
2.2.1.2.2. Bộ lọc quang học

Đối với các Video thu được từ hệ thống quang học kết hợp với hệ thống cảm biến, bộ lọc này có mục đích duy nhất là chuyển đổi ánh sáng và thông tin quang học thành dữ liệu kỹ thuật số dưới dạng một luồng video. Rất nhiều thông tin liên quan trực tiếp đến ánh sáng và thông tin quang học ban đầu được thiết bị thu nhận được ẩn trong cấu trúc của video. Mục đích của bộ lọc quang học là để trích xuất thông tin này cho phép người điều tra tìm kiếm sự bất thường trong các mẫu thông tin quang học. Những điểm bất thường này có liên quan trực tiếp đến vật lý quang học. Do đó, cần phải có một số kiến thức về những hiện tượng này để giải thích chính xác kết quả.

• Bộ lọc **Fluor** được sử dụng để nghiên cứu màu sắc của hình ảnh bất kể mức độ chói của nó. Bộ lọc tạo ra một hình ảnh bình thường trong đó màu sắc của hình ảnh ban đầu đã được khôi phục độc lập với độ chói. Biến đổi cơ bản theo công thức:

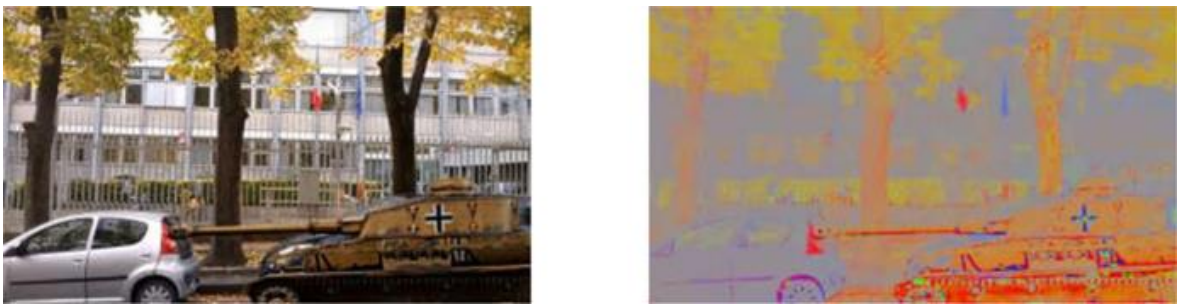
$$\begin{pmatrix} \text{red} \\ \text{green} \\ \text{blue} \end{pmatrix} = \begin{pmatrix} \frac{\text{red}}{\text{red}+\text{green}+\text{blue}} \\ \frac{\text{green}}{\text{red}+\text{green}+\text{blue}} \\ \frac{\text{blue}}{\text{red}+\text{green}+\text{blue}} \end{pmatrix}$$

Hình 2.4. Bộ chuyển đổi màu của bộ lọc Flour



Hình 2.5. Nguyên tắc chiếu được thực hiện bởi bộ lọc Fluor

Như được minh họa trong Hình 2.5, ở dạng 2D hoặc 3D, các điểm ảnh màu có các thành phần Đỏ, Xanh lục và Xanh lam được chiếu lên hình cầu có tâm là màu đen sao cho chuẩn của vectơ mới (đỏ, lục, lam) luôn bằng 1. Trên hình ảnh 2D, các điểm màu đen đại diện cho các màu khác nhau nhưng hình chiếu của chúng trên cung tròn nằm trong cùng một vùng tạo ra cùng một màu sắc của hình ảnh Fluor. Mặt khác, các điểm ảnh tối, được vẽ dưới dạng các điểm có màu xám trong hình ảnh nhưng thực tế có thể có màu sắc khác và hình chiếu của chúng trên cung sẽ tăng cường những khác biệt này và có thể cho phép người dùng phân biệt giữa chúng. Quá trình chuẩn hóa này được thực hiện bởi bộ lọc Fluor giúp nó có thể phá vỡ sự tương đồng của màu sắc khi nó được hệ thống thị giác của con người cảm nhận và



Hình 2.6. Đầu ra của bộ lọc Fluor trên video xe tăng đã bị chỉnh sửa.

làm nổi bật các màu có sự khác biệt rõ rệt hơn dựa trên màu sắc thực tế của chúng.

Hình 2.7. Đầu ra của bộ lọc Focus trên video xe tăng đã bị chỉnh sửa.

• Bộ lọc **Focus** được sử dụng để xác định các khu vực sắc nét trong một hình ảnh hoặc các vùng có độ nét mạnh hơn. Khi một hình ảnh sắc nét, nó có đặc tính chứa chuyển đổi bất thường trái ngược với mức độ thay đổi của màu sắc ở ranh giới của một đối tượng. Một hình ảnh có độ sắc nét cao chứa lượng tần số cao nhiều hơn, trong khi ngược lại, các tần số cao không phù hợp khi đối tượng bị mờ hoặc mất nét. Ước tính độ sắc nét này được bộ lọc Focus thực hiện thông qua phép biến đổi wavelet [26]. Bộ lọc Focus chỉ xem xét độ phân giải của wavelet thông qua bộ lọc phi tuyến dựa trên việc xử lý bộ RGB của mỗi khung. Nó tạo ra bố cục màu sai nơi mà các vùng tần số thấp bị mờ vẫn có màu xám và các đường viền sắc nét xuất hiện màu.

• Bộ lọc **Acutance** đề cập đến thuật ngữ vật lý để chỉ độ sắc nét trong hình ảnh. Thông thường, nó là một phép đo đơn giản về độ dốc của gradient cục bộ nhưng ở đây nó được chuẩn hóa với giá trị cục bộ của các mức xám, khác với bộ lọc Focus. Bộ lọc Acutance được tính bằng tỷ số giữa các đầu ra của một bộ lọc high-pass và một bộ lọc low-pass. Trong thực tế, có thể sử dụng hai bộ lọc Gaussian với các kích thước khác nhau. Do đó, phương trình sau đặc trưng cho tiến trình lọc

Hình 2.8. Đầu ra của bộ lọc Acutance trên video xe tăng đã bị chỉnh sửa.

Acutance:

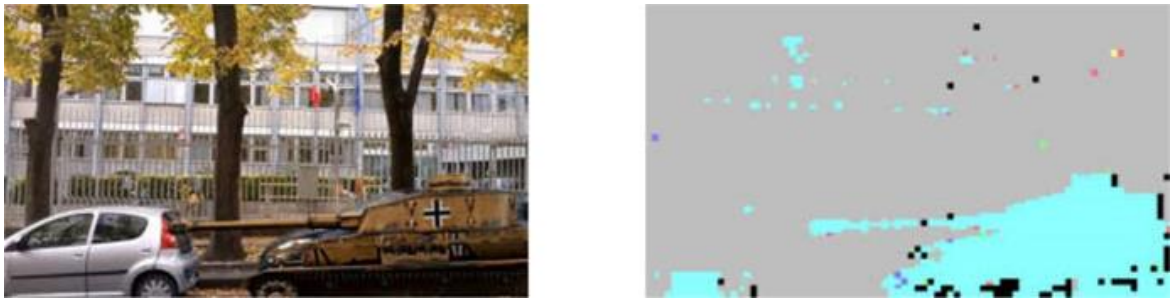
2.2.1.2.3. Các bộ lọc thời gian

Những bộ lọc này nhằm mục đích làm nổi bật hoạt động của luồng video theo thời gian. Việc nén video MPEG-4 khai thác khả năng dư thừa theo thời gian để giảm kích thước video nén. Đây là lý do tại sao nén một video phức tạp hơn nhiều so với nén một chuỗi các hình ảnh. Hơn nữa, trong nhiều khung hình, MPEG-4 kết hợp các dự đoán theo một hướng hoặc theo chiến lược tiến/lùi, do đó, việc biểu diễn khung phụ thuộc nhiều vào nội dung khung và mức độ lượng tử hóa. Do đó, việc phân tích hành vi theo thời gian của các tham số lượng tử hóa có thể giúp chúng ta phát hiện ra sự mâu thuẫn trong biểu diễn khung.

• Bộ lọc **Cobalt** so sánh video gốc với phiên bản sửa đổi được định dạng lại bằng MPEG-4 với mức chất lượng khác (tốc độ bit tương ứng khác). Nguyên tắc của bộ lọc Cobalt rất đơn giản: một bộ quan sát video về các lỗi giữa video đầu tiên và video được định lượng lại bằngMPEG-4 với mức chất lượng thay đổi hoặc mức tốc độ bit thay đổi, nếu mức lượng tử hóa trùng với mức chất lượng thực sự được sử dụng trên khu vực sửa đổi nhỏ, thì sẽ không có lỗi ngay tại đó. Thực hành này khá giống với thuật toán JPEG Ghosts, trong đó ảnh JPEG được giải nén lại và ảnh mới được trừ khỏi ảnh gốc, để làm nổi bật cục bộ các điểm không nhất quán (“bóng mờ”) tương ứng với các đối tượng được thêm vào từ các ảnh có chất lượng khác nhau (thuật toán ELA cũng theo một cách tiếp cận tương tự).

Hình 2.10. Đầu ra của bộ lọc Cobalt

• Bộ lọc các **Vectơ chuyển động** mang lại sự trình bày dựa trên màu sắc của các khối chuyển động khi được mã hóa vào luồng video. Thông thường, loại biểu diễn này sử dụng các mũi tên để hiển thị các sự dịch chuyển của các khối. Điều đáng chú ý là hệ thống mã hóa không nhận ra 'đối tượng' mà chỉ xử lý các khối. Các vec-tơ chuyển động được mã hóa trong luồng video để tái tạo lại tất cả các khung không phải là khung chính (nghĩa là không phải khung được mã hóa nội bộ mà là các khung được mã hóa liên khung bằng cách sử dụng thông tin từ các khung khác). Sau đó, một đối tượng của cảnh có một tập hợp các vectơ chuyển động được liên kết với mỗi khối macro bên trong nó. Những chuyển động này được đại diện bởi bộ lọc Vectơ chuyển động phải đồng nhất và nhất quán, nếu không có khả năng cao là



Hình 2.11. Đầu ra của bộ lọc vector chuyển động

một số hoạt động đáng ngờ liên quan việc chỉnh, sửa video đã xảy ra.

• Bộ lọc **Temporal** được sử dụng để áp dụng chuyển đổi thời gian trên video, chẳng hạn như làm mịn hoặc điều chỉnh thời gian. Nó cũng được sử dụng để so sánh khung hình chỉ tập trung vào sự phát triển của độ sáng theo thời gian. Bộ lọc Temporal được tính là sự khác biệt giữa khung hình với khung hình theo thời gian như được nêu trong phương trình sau:

$$frame_{\text{Temporal Filter}}(t) = frame(t) - frame(t-1) \qquad (2.2)$$

Công thức (2.2) được áp dụng trên mỗi kênh màu của các khung hình để đầu

Hình 2.12. Đầu ra của bộ lọc Temporal

ra của bộ lọc cũng là hình ảnh màu.

Nhận xét: Phương pháp tiếp cận dựa trên đặc trưng luồng đa phương tiện là một phương pháp tiếp cận hiệu quả trong khi giám định video ở mức khung hình vẫn là chiến lược khả thi duy nhất với dữ liệu có sẵn hạn chế. Trong đó, bộ lọc Q4 và bộ lọc Cobalt thường được ứng dụng thực tế nhiều hơn trong việc phân loại dữ liệu đầu ra của các hệ thống giám định video bởi tính trực quan và hiệu quả của nó.

2.2.1.3. Phương pháp tiếp cận dựa trên tín hiệu audio

Hiện nay, âm thanh đã trở thành một phần không thể thiếu trong các video, nó là một trong những thành phần lưu trữ và truyền tải thông tin, khiến cho video trở thành một nội dung đa phương tiện phổ biến có khả năng ghi lại các sự kiện và truyền đạt thông điệp sinh động, cụ thể trên khắp thế giới. Cùng với sự phát triển của khoa học kỹ thuật hiện đại, việc sửa đổi hoặc thay thế âm thanh hoặc thay thế video thường khá dễ thực hiện và có thể thay đổi nội dung thông tin gốc một cách đáng kể; trong đó, so với việc thay đổi nội dung video thì việc thay đổi nội dung âm thanh thường rất khó để con người có thể phát hiện. Tuy nhiên, những sửa đổi này thường để lại dấu vết bởi sự khác biệt giữa hình ảnh và âm thanh của video - một đặc trưng chúng ta có thể khai thác nhằm phát hiện sự chỉnh sửa, cắt/ghép video. *Phát hiện sự không nhất quán về âm thanh - hình ảnh* là một hướng tiếp cận mà nhiều nhà khoa học đang nghiên cứu, phát triển để phát hiện và mô tả đặc điểm của nhiều loại mâu thuẫn liên quan đến các khía cạnh khác nhau tạo nên một video hoàn chỉnh [31]. Ví dụ, một video quay cảnh một người đang nói chuyện hoặc một sự kiện nào đó đang diễn ra; nó được quay trong một môi trường nhất định, chẳng hạn như trong một căn phòng nhỏ hoặc ngoài trời, trên một con phố đông đúc… người đang nói có thể là một cá nhân nổi tiếng hoặc bất cứ ai; khi người đó nói, âm thanh giọng nói và chuyển động môi của họ tạo ra những tín hiệu sóng âm liên quan đến những gì họ đang nói và vị trí đầu của họ có thể di chuyển so với micrô theo những cách ảnh hưởng đến âm thanh; ngoài ra, trong thực tế, có thể có các hoạt động khác và âm thanh có thể dự đoán, có thể được phân loại rõ ràng (ví dụ: đám đông, nước

chảy…) hoặc tạo ra các dấu hiệu liên kết chặt chẽ hình ảnh và âm thanh. Trong cách tiếp cận này, các kênh âm thanh và hình ảnh phải khớp nhau, trừ khi video đó đã bị chỉnh sửa theo một cách nào đó. Việc phát hiện ra nhiều bất đồng là một dấu hiệu chứng minh của một số loại chỉnh sửa video.

Sự không nhất quán của *môi trường xung quanh* và *đặc trưng âm thanh/hình ảnh người nói* đại diện cho các loại chỉnh sửa khác nhau và là bằng chứng xác thực, cũng như thách thức cho công tác giám định video. Một hệ thống cơ bản để triển khai phát hiện video đã bị chỉnh sửa theo hướng tiếp cận này thường gồm 3 giai đoạn chính: (1) Phát hiện sự mâu thuẫn âm thanh và hình ảnh trong các cảnh; (2) Phát hiện sự không nhất quán trong những thay đổi của âm thanh và nhận dạng hình ảnh trong video; và (3) Xây dựng các bộ dữ liệu và triển khai để đánh giá hiệu quả thuật toán.

Một số kỹ thuật trong phương pháp tiếp cận này như:

(1) Phân tích, phát hiện âm thanh của cảnh: Hệ thống phát hiện âm thanh cảnh dựa trên một I-vectơ được mô hình hóa bằng cách sử dụng một nền tảng Gaussian. Số lượng của bộ Gaussian dựa trên số lượng các lớp. Phương pháp trích xuất các đặc trưng giọng nói (Mel Frequency Cepstral Coefficients - MFCC) của 20 chiều được sử dụng để trích xuất đặc điểm âm thanh từ các cửa sổ (windows) 25ms cứ sau 10ms. Nội dung được cung cấp với kết quả *deltas* và *double deltas* trong các đặc trưng 60 chiều. Công cụ trích xuất I-vectơ sử dụng Mô hình nền phổ quát Gaussian 1024 (UBM); I-vectơ đã được dự báo đến 200 chiều bằng cách sử dụng Phân tích phân biệt tuyến tính (LDA). Cuối cùng, một mô hình Gaussian được ước lượng với các i-vectơ thuộc mỗi lớp; các mô hình phát hiện âm thanh cảnh có thể được huấn luyện trên một tập hợp con của bộ dữ liệu Placing chứa khoảng 600 video, là một tập hợp con của bộ dữ liệu YFCC100M [17].

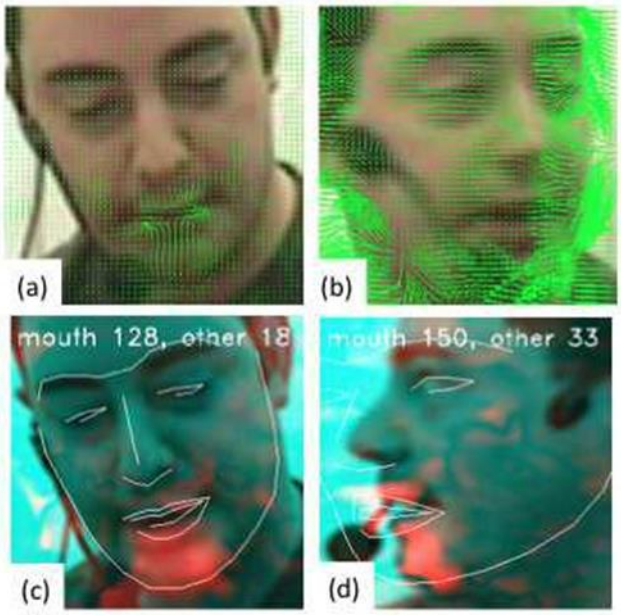
(2) Phát hiện cảnh không nhất quán: Trong [31], các nhà nghiên cứu đã xây dựng hệ thống, mà ở đó thông qua phát triển các tập lệnh Python, họ có thể để cắt và dán các track âm thanh và hình ảnh lại với nhau để tạo thành các video mới và

chú thích từ các video nguồn có chú thích. Một ví dụ tương đối đơn giản về điều này là trích xuất dữ liệu âm thanh và hình ảnh từ một clip được gắn nhãn là sa mạc và thay thế một nửa âm thanh bằng âm thanh từ một clip được gắn nhãn là cuộc họp trong nhà. Sau quá trình phân tích, chúng ta có thể dễ dàng nhận thấy sự mâu thuẫn giữa âm thanh và hình ảnh trong video.

(3) Phát hiện sự không nhất quán của người nói: Một số sự chỉnh sửa liên quan đến việc thay thế lời nói của một người bằng lời nói của người khác, chẳng hạn như lồng tiếng hoặc có thể thay thế khuôn mặt (deep-fakes) nhưng không chỉnh sửa âm thanh. Những sự chỉnh sửa này có thể tạo ra sự mâu thuẫn thường có thể quan sát được giữa trạng thái không hoạt động của người nói hay cảm xúc của khuôn mặt đang nói chuyện một cách tương đối rõ ràng. Dựa trên cơ sở dữ liệu về các đối tượng đã biết với các mẫu giọng nói và khuôn mặt của họ, chúng ta có thể kiểm tra các kết quả nghe-nhìn không nhất quán với cơ sở dữ liệu. Theo đó, để phát hiện cần tìm ít nhất hai phân đoạn (khoảng thời gian) mà nhận dạng âm thanh hoặc hình ảnh của người nói khác nhau, bằng cách tính toán từng phương thức một cách độc lập trước và sau đó kiểm tra sự mâu thuẫn giữa chúng. Trong mỗi phương thức, phát hiện các phân đoạn giọng nói và ước tính khả năng người nói giống nhau trong mỗi cặp phân đoạn này. Xác suất không nhất quán sau đó được cộng lại cho các cặp phân đoạn âm thanh và hình ảnh chéo nhau.

- *Phân tích âm thanh, phân cực người nói:* Phân cực âm thanh người nói là quá trình phân chia audio thành các phân đoạn đồng nhất với âm thanh người nói và tổ chức các phân đoạn thành các cụm có cùng đặc điểm người nói. Sự phân cực ở đây dựa trên các i-vectơ được phân cụm bằng cách sử dụng hệ thống nhận dạng âm thanh người nói theo i-vectơ PLDA [23]. Hệ thống nhận dạng âm thanh qua khẩu hình người nói để tạo ra một ma trận điểm cho các i-vectơ từ âm thanh được phân cực. Các giá trị này sau đó được chuyển thành ma trận khoảng cách phân đoạn đến âm thanh người nói thu được và vị trí của các phân đoạn trong video để tìm kiếm sự không nhất quán liên quan đến hình ảnh đầu ra.

- *Phân tích hình ảnh nhằm phát hiện và theo dõi các khuôn mặt đang nói chuyện:* Để phân tích tính nhất quán về mặt nghe nhìn của người nói, cần tìm các phân đoạn trong video có sự xuất hiện của người đang nói chuyện và xác định từng



Hình 2.13. Phát hiện người nói sử dụng luồng quang học

khuôn mặt đó.

Người nói được phát hiện bằng cách sử dụng kết hợp các điểm mốc trên khuôn mặt và luồng quang học (có thể sử dụng các thư viện được tích hợp sẵn trong thư viện OpenCV). Hình 2.13 hiển thị các ví dụ về luồng quang học và các điểm mốc được căn chỉnh cho các mặt AMI. Kết quả thử nghiệm trên dữ liệu AMI cho thấy rằng việc phát hiện giọng nói bằng cách sử dụng những thay đổi trong các mốc khuôn mặt được căn chỉnh, chẳng hạn như môi trên và môi dưới, không phải lúc nào cũng hoạt động tốt, đặc biệt là ở các tư thế khuôn mặt khó đối với các phương pháp căn chỉnh hiện tại [31].

- *Phát hiện sự không nhất quán của người nói:* Hình 2.14 hiển thị một phân đoạn giọng nói và âm thanh dựa trên khẩu hình của người nói. Mỗi hàng màu xanh lá cây là một phần âm thanh khuôn mặt đang nói và mỗi hàng màu đỏ là một âm thanh của video. Thời gian là theo chiều ngang và các dấu tích ở giữa là các giây.

Hình 2.14. Âm thanh của khẩu hình và âm thanh video

(a) và (b) là tương thích; (c) và (d) là có sự mâu thuẫn.

Nhận xét: Phương pháp tiếp cận này có thể triển khai theo hai hướng phát hiện video giả mạo bằng cách khai thác hai loại nội dung nghe-nhìn, gồm: loại cảnh và xác thực âm thanh người nói. Các yếu tố mới trong cách tiếp cận này bao gồm phương pháp xây dựng bản đồ ngữ nghĩa giữa các tập hợp cảnh âm thanh và hình ảnh không khớp, xây dựng hệ thống có khả năng phát hiện sự không nhất quán về mặt nghe nhìn từ các track âm thanh dựa trên hình ảnh khuôn mặt và âm thanh thực của video tương ứng. Việc thử nghiệm trên một bộ sưu tập video giả mạo đã cho thấy nhiều kết quả hứa hẹn cho phương pháp này. Tuy nhiên, phương pháp tiếp cận này mới dừng ở việc phát hiện các video bị giả mạo, chưa phát hiện được vị trí, điểm cắt, ghép trong video theo yêu cầu Đề tài đặt ra.

2.2.2. Đề xuất thuật toán giải quyết bài toán

Mặc dù hiện nay, một số nghiên cứu đã được đề xuất để đối phó với việc phát hiện video giả mạo nói chung và các điểm cắt, ghép trong đó nói riêng với kết quả đầy hứa hẹn nhưng vẫn tồn tại nhiều vấn đề gây tranh cãi, như:

- Vì việc phát hiện giả mạo hình ảnh là một vấn đề khó, thời gian tính toán tương đối cao đối với các kỹ thuật dựa trên khối (block-based), vì tất cả các pixel hoặc các tính năng trích xuất phải được kiểm tra trên mỗi khối, đôi khi, sự giả mạo trên quy mô lớn không thể phát hiện được. Ngược lại, các kỹ thuật dựa trên đặc trưng sử dụng các thuật toán có độ phức tạp thấp hơn. Do đó, việc giải quyết tính cân bằng giữa tốc độ và độ chính xác hiện là một vấn đề đầy thách thức.

- Việc xác định thuật toán trích xuất đặc trưng tối ưu nhất không dễ dàng và kết quả cuối cùng phụ thuộc nhiều vào kỹ thuật được sử dụng, có thể là kỹ thuật dựa trên keypoints hoặc dựa trên khối.

- Trong nhiều trường hợp, các phương pháp hiện có không phát hiện được các vùng trùng lặp đặc biệt nhỏ (gây ra bởi thao tác copy-move), do đó độ chính xác sẽ rất thấp. Một số phương pháp không xác định được nhiều vùng trùng lặp, các kỹ thuật dựa trên keypoints không thể giải quyết việc giả mạo đã được làm mịn.

- Đôi khi, các phương pháp dựa trên khối chính xác hơn các phương pháp dựa trên keypoint trong việc xác định hình dạng của các vùng trùng lặp.

- Cả các tính năng chính và đối sánh dựa trên khối đều gặp khó khăn trong việc phát hiện chính xác vùng hình dạng được làm mịn.

- Hầu hết các phương pháp phát hiện nhân bản không thể phát hiện các loại tấn công khác nhau cùng một lúc (ví dụ: nén, chia tỷ lệ và thêm nhiễu, di chuyển sao chép và nhân bản).

- Quá khó khi chỉ ứng dụng một phương pháp hoặc thuật toán phát hiện giả mạo hình ảnh duy nhất để có thể phát hiện toàn bộ hình ảnh giả mạo. Vì vậy, cần sự kết hợp của nhiều phương pháp liên quan.

Đáng chú ý, những nghiên cứu, khảo sát gần đây đã cho thấy, tương ứng với mỗi loại giả mạo trong video sẽ có những kỹ thuật phát hiện phù hợp [27], cụ thể:

Bảng 2.1. Các kỹ thuật phát hiện giả mạo video

Loại giả mạo	Kỹ thuật phát hiện giả mạo
Sao chép-di chuyển (nhân bản)	Khớp khối (Block matching), Biến đổi Cosine rời rạc (DCT), Phân tích thành phần chính (PCA), Tương quan chuỗi
Nối (Splicing)	Phân tích quang phổ hai mặt, phân tích kết hợp song song, ước tính biến thiên nhiễu, thống kê bậc cao.
Re-sampling	Phương pháp thống kê (thuật toán EM)
Nén JPEG kép	Ước tính nén JPEG (phân tích tần số)
Chỉnh sửa (độ sáng, độ nhiễu phi tuyến tính)	Thuật toán EM, thống kê đơn bậc cao
Tăng cường đa phương tiện	Công cụ ước tính thống kê mù (công cụ ước tính mờ, ước lượng nhiễu, ước tính biến đổi hình học)
Biến đổi hình học (dịch, xoay, chia tỷ lệ, nghiêng, phản chiếu)	Cung cấp thông tin không gian giữa các khối được sao chép và đối tượng liền kề
Xử lý hậu kỳ (nén JPEG / MPEG, nhiễu, nhòe)	Loại bỏ bất kỳ dấu hiệu thao túng đáng chú ý nào đặc biệt là các cạnh sắc nét

Do đó, dựa trên những nghiên cứu hiện có này, học viên định hướng triển khai thực nghiệm bằng cách nghiên cứu xây dựng một hệ thống phần mềm giám định video theo ***phương pháp tiếp cận dựa trên kỹ thuật biến đổi Cosin rời rạc*** - một phương pháp phổ biến trong phát hiện video giả mạo loại copy-move (loại giả

mạo thường gặp hiện nay) và được đánh giá có kết quả với độ chính xác cao; hướng đến mục tiêu làm nổi bật các dấu vết nào do giả mạo video để lại, với trọng tâm là xác định sự gián đoạn trong các khía cạnh thời gian của video làm nền tảng cho các nghiên cứu phát triển tiếp theo trong tương lai; qua đó đánh giá hiệu quả của phương pháp thực nghiệm trên. Đối với nhiều lĩnh vực phân tích dữ liệu khác, như mạng nơ ron sâu cũng cho kết quả rất hứa hẹn trong việc phát hiện giả mạo, trong đó có các video hoặc ảnh kỹ thuật số. Vì vậy, với sự phát triển của một số công cụ phân tích nhằm cung cấp cho người dùng phương tiện có khả năng làm nổi bật những điểm không nhất quán trong nội dung video, học viên định hướng sau khi xây dựng hệ thống phát hiện điểm cắt, ghép trong video sẽ tiến hành nghiên cứu, phát triển một phương pháp học sâu nhằm phân tích kết quả đầu ra của các công cụ giám định này và tự động phát hiện các video giả mạo.

Qua quá trình khảo sát về thực trạng thế giới công nghệ hiện nay liên quan đến mức độ phù hợp của hướng giải quyết bài toán trên, công cụ giám định mà học viên phát triển có tiềm năng rất lớn trong việc xử lý cục bộ các tập dữ liệu video giả mạo, cũng như phương pháp học sâu mà học viên nghiên cứu đã góp phần tích cực và ảnh hưởng rất lớn tới hiệu quả tự động phát hiện giả mạo video. Dựa trên kết quả thử nghiệm đối với dữ liệu điểm chuẩn và thế giới thực, đồng thời phân tích kết quả, học viên nhận thấy rằng phương pháp được đề xuất mang lại kết quả đầy hứa hẹn so với phương pháp hiện đại, đặc biệt là đối với khả năng khái quát hóa của thuật toán đối với dữ liệu chưa biết được lấy từ thế giới thực.

Với những thách thức này, học viên bắt đầu nghiên cứu xây dựng, triển khai thành phần giám định video nhằm hướng tới phát triển một hệ thống có thể hỗ trợ các chuyên gia trong việc xác định các video giả mạo hoặc nâng cao hiện đại hóa lĩnh vực kỹ thuật hình sự theo hướng này. Bắt đầu bằng cách nghiên cứu trong giám định hình ảnh và kết hợp chuyên môn, học viên mở rộng nghiên cứu một số thuật toán, còn được gọi là "bộ lọc", nhằm xử lý video và giúp người dùng cụ thể hóa các điểm mâu thuẫn đáng ngờ trong video. Những bộ lọc này hướng tới khả năng đưa ra kết quả được hiển thị cho người dùng, giúp họ xác minh video một cách trực quan

là cơ sở để nghiên cứu, xây dựng hệ thống tự động hóa quá trình phát hiện bằng cách đào tạo một kiến trúc mạng nơ-ron nhân tạo để phát hiện những điểm không nhất quán này và phân loại video là xác thực hoặc bị giả mạo.

Học viên tập trung vào phát hiện giả mạo video và không đề cập đến các hình thức xác minh khác, chẳng hạn như phân tích thành phần chính nội dung video hoặc xem xét siêu dữ liệu hoặc thông tin theo ngữ cảnh. Nó được dành riêng cho các phương tiện được sử dụng để theo dõi các dấu vết yếu (hoặc chữ ký) để lại bởi quá trình giả mạo trong nội dung video được mã hóa. Nó giải thích cho tính toàn vẹn của mã hóa, không gian, thời gian, màu sắc và sự liên kết lượng tử hóa. Hai cách tiếp cận bổ sung được trình bày, một phương pháp xử lý nội địa hóa giả mạo, tức là sử dụng các bộ lọc để tạo ra các bản đồ đầu ra nhằm mục đích làm nổi bật vị trí hình ảnh có thể đã bị giả mạo và được thiết kế để người dùng hiểu được và một phương pháp xử lý phát hiện giả mạo, nhằm mục đích tạo ra đầu ra một giá trị cho mỗi video cho biết xác suất video đó bị giả mạo.

Chương 3 - THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ

Đề tài phát hiện video bị chỉnh sửa hiện nay vẫn là một trong những vấn đề khó, thách thức các chuyên gia, khi phải đối mặt với ngày càng nhiều kỹ thuật chỉnh sửa video hiện đại, tinh vi. Do đó đề tài phát hiện điểm cắt, ghép trong video lại càng khó hơn nữa, đây là giai đoạn tiếp theo của hệ thống sau khi đã nhận diện được các video bị chỉnh sửa; một số phương pháp hay nghiên cứu hiện nay mới chỉ dừng lại ở việc phát hiện các video bị chỉnh sửa, chưa thể phát hiện được các vị trí hoặc loại chỉnh sửa mà video đã bị tác động. Qua quá trình nghiên cứu các phương pháp sử dụng trong phát hiện điểm cắt ghép trong video, học viên đã tiến hành xây dựng một chương trình dựa trên thuật toán bộ lọc Cosin rời rạc để thực nghiệm và đánh giá về các phương pháp nghiên cứu phát hiện điểm cắt, ghép trong video, làm tiền đề phục vụ cho các nghiên cứu sau này.

3.1. Giới thiệu chương trình

3.1.1. Nền tảng công nghệ

- Chương trình được xây dựng trên nền tảng ngôn ngữ lập trình bậc cao Python version 3, sử dụng công cụ lập trình Pycharm - môi trường phát triển hoàn hảo giành cho ngôn ngữ lập trình Python để thực hiện. Các thư viện hỗ trợ bao gồm:

+ Thư viện OpenCV là thư viện rất mạnh trong thực hiện các thao tác xử lý ảnh trong Python nói riêng và các ngôn ngữ lập trình khác nói chung.

+ Thư viện Scipy là thư viện hỗ trợ các thuật toán liên quan xử lý học máy mà ta sẽ sử dụng trong xử lý dữ liệu, tìm ra những điểm cắt ghép trong video.

+ Thư viện Pillow là thư viện cũng rất mạnh để xử lý ảnh trong Python

+ Thư viện Numpy là thư viện cốt lõi phục vụ cho khoa học máy tính, hỗ trợ tính toán trên các mảng nhiều chiều, kích thước lớn mà ta sử dụng trong chương trình.

- Chương trình sử dụng phương pháp biến đổi Cosin rời rạc trong xử lý ảnh kết hợp cùng với phương pháp học máy để tìm kiếm những vùng bị cắt ghép trong

mỗi frame ảnh của video đã trích xuất. Kết quả thu được góp phần đánh giá hiệu quả của chương trình phát hiện điểm cắt ghép trong video sử dụng phương pháp biến đổi Cosin rời rạc, những frame ảnh đã được khoanh vùng chỉnh sửa, làm cơ sở đưa ra nhận định những video đã bị thay đổi nội dung.

3.1.2. Nguồn dữ liệu

Việc đánh giá một phương pháp có hiệu quả hay không thì lựa chọn dữ liệu đầu vào là rất quan trọng. Qua nghiên cứu, học viên đã lựa chọn bộ dữ liệu của InVID Fake Video Corpus, được phát triển trên nguồn của dự án InVID. Fake Video Corpus (FVC) bao gồm rất nhiều video đã được chỉnh sửa, cắt ghép nội dung và bên cạnh đó là số lượng lớn video gốc không bị chỉnh sửa. Nguồn video được tải lên các nền tảng mạng xã hội gồm cả Youtube và Facebook. Các video sẽ được xử lý trước khi đưa vào mô hình thử nghiệm theo chuẩn mã hóa H.264/AVC, định dạng file *.mp4, độ dài mỗi video khoảng từ 10-30s.

Để thử nghiệm và đánh giá các phương pháp đã nghiên cứu trong những phần trước, học viên sẽ lựa chọn ngẫu nhiên ra các video trong tập dữ liệu video để thử nghiệm, trong đó sẽ sử dụng các video có chỉnh sửa nhằm áp dụng các thuật toán đã đưa ra để đánh giá hiệu quả của phương pháp áp dụng.

Từ tập dữ liệu đã được chọn, học viên đã chọn lọc lấy ra 5 video có nội dung đã bị chỉnh sửa. Sau đó, sử dụng phần mềm chỉnh sửa video để chuẩn hóa kích thước khung hình của video để làm dữ liệu đầu vào cho việc thử nghiệm chương trình.

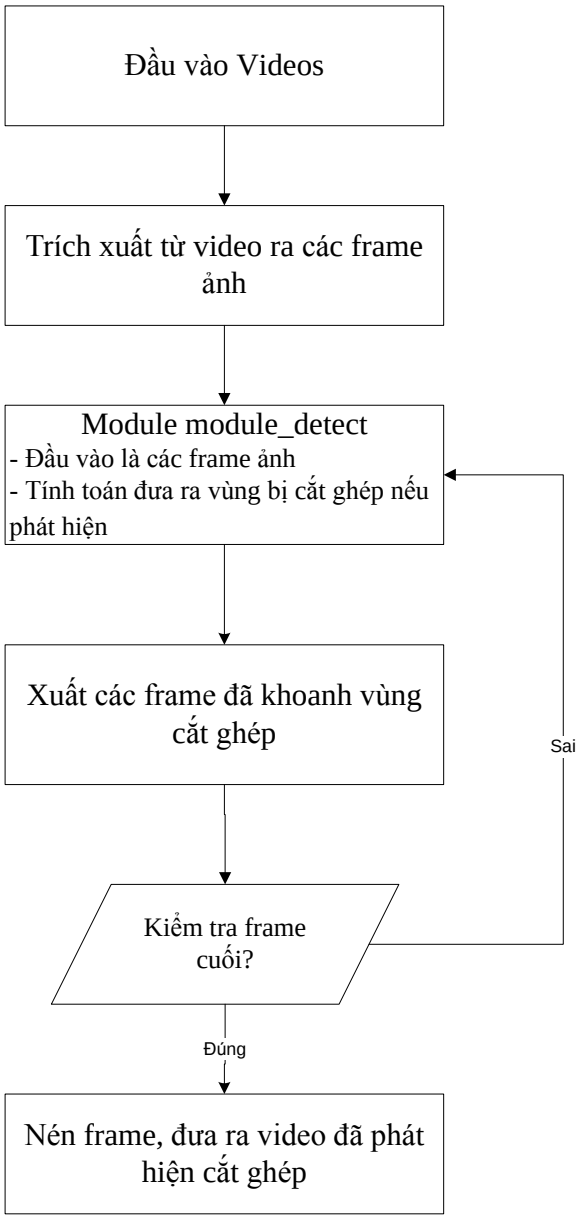
3.2. Cấu trúc chương trình

Chương trình phát hiện điểm cắt ghép trong video được nghiên cứu và xây dựng trên nền tảng ngôn ngữ lập trình Python, do tính ưu việt của ngôn ngữ cũng như sự hỗ trợ mạnh mẽ bởi bộ thư viện đồ sồ, cộng đồng đông đảo người lập trình trên toàn thế giới. Trong chương trình thử nghiệm, học viên chia ra các module để thực hiện bao gồm:

- ***Module xử lý dữ liệu đầu vào:*** Module này thực hiện việc đọc dữ liệu video đầu vào, trích xuất video thành các frame ảnh dưới dạng ảnh xám và lưu lại vào trong từng thư mục riêng biệt. Thư viện của ngôn ngữ lập trình Python là OpenCV hỗ trợ rất mạnh trong xử lý công việc này.

- ***Module xử lý phát hiện điểm cắt ghép:*** Đây là module quan trọng nhất, đảm nhận nhiệm vụ chính đó là xử lý hình ảnh. Áp dụng phép biến đổi Cosin rời rạc kết hợp với các thuật toán học máy trong bộ thư viện của Python để tìm ra các điểm bất thường, có khả năng là những điểm bị chỉnh sửa trong mỗi khung ảnh của video đã trích xuất. Sau đó xuất khung ảnh đã qua xử lý ra thư mục đầu ra riêng biệt.

- ***Module chuyển đổi ảnh sang video:*** Sau khi đã xử lý toàn bộ các khung ảnh, công việc tiếp theo là phải chuyển đổi các khung ảnh đã xử lý thành video kết quả hoàn chỉnh. Các video được chuyển đổi theo các chuẩn mà thư viện Python hỗ trợ.



Hình 3.1. Cấu trúc chương trình

Theo Hình 3.1, thứ tự thực hiện các bước được thực hiện như sau:

- Bước 1: Dữ liệu đầu vào là các video được chuẩn hóa dưới dạng MPEG, điều chỉnh theo một tỉ lệ khung hình nhất định, cụ thể là 320x320 pixel.

- Bước 2: Trích xuất các frame từ video đầu vào.

- Bước 3: Đưa các hình ảnh vào trong Module phát hiện cắt ghép đã lập trình sẵn.

- Bước 4: Xuất frame ảnh đã được kiểm tra trong Module Detect trong bước 3.

- Bước 5: Kiểm tra frame đã là frame cuối chưa. Nếu đúng thì chuyển sang bước 6, nếu sai thì quay lại bước 3.

- Bước 6: Nén các frame đã xuất tại bước 4 thành video hoàn chỉnh.

3.2.1. Xử lý dữ liệu đầu vào

Video ngày nay có rất nhiều định dạng, phổ thông như định dạng MP4, AVI, FLV… Trong chương trình, ta sử dụng dữ liệu đầu vào định dạng *.MP4. Mỗi video có một kích thước khung hình, độ lớn khác nhau, nó sẽ ảnh hưởng tới hiệu suất xử lý việc phát hiện điểm cắt ghép trong video. Qua thử nghiệm, với mỗi khung hình khác nhau sẽ có thời gian xử lý khác nhau, cụ thể theo Bảng 3.1. Thời gian xử lý tương ứng với kích thước khung hình dưới đây:

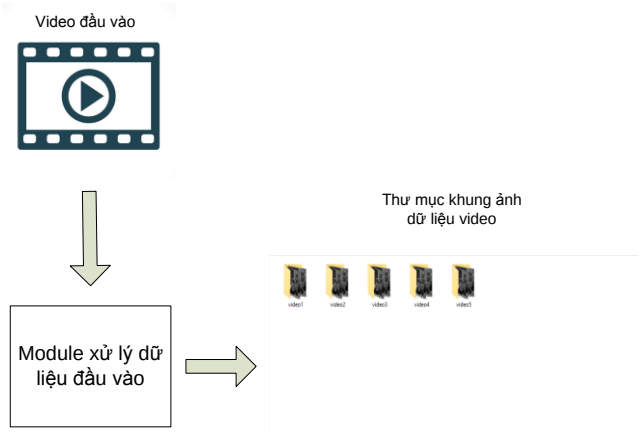
Bảng 3.1. Thời gian xử lý tương ứng với kích thước khung hình

Tỉ lệ khung hình	Thời gian xử lý trung bình
1024 x 768	240s
256 x 256	14s
512 x 512	60s
360 x 360	40s

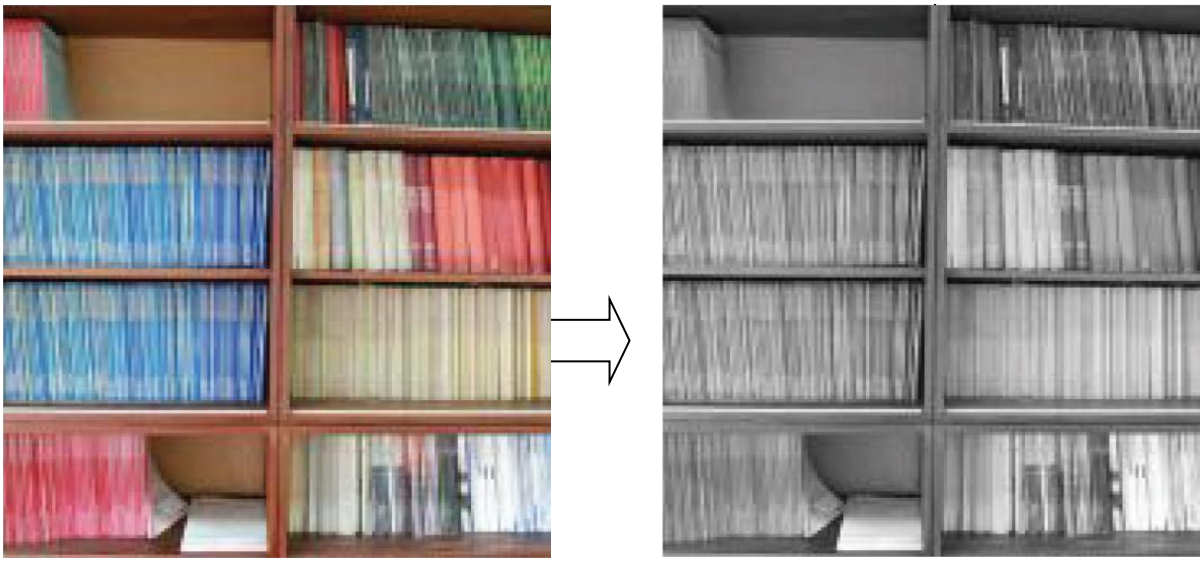
Thời gian xử lý ở trên còn phụ thuộc rất nhiều vào yếu tố cấu hình phần cứng, độ sắc nét của video. Trong chương trình này, học viên xin áp dụng thử nghiệm trên video có độ phân giải 360 pixels và chuẩn hóa sang video vuông có kích thước dài, rộng như nhau (bằng 360 pixel) để thuận lợi cho việc thử nghiệm trên mô hình.

Sau khi đã chỉnh sửa video, học viên tiến hành tách các khung hình từ video đã chọn và chỉnh sửa. Đồng thời, trong quá trình tách các khung hình từ video, ta sẽ tiến hành chuyển đổi các khung hình sang dạng ảnh xám để làm đâu vào cho quá trình xử lý ở module tiếp theo. Mỗi khung hình sẽ được lưu lại dưới định dạng ảnh *.png tại thư mục cố định. Module này được xây dựng trên cơ sở áp dụng các hàm

trong thư viện OpenCV của Python - một thư viện rất mạnh và quen thuộc trong xử lý đa phương tiện.



Hình 3.2. Xử lý dữ liệu đầu vào video



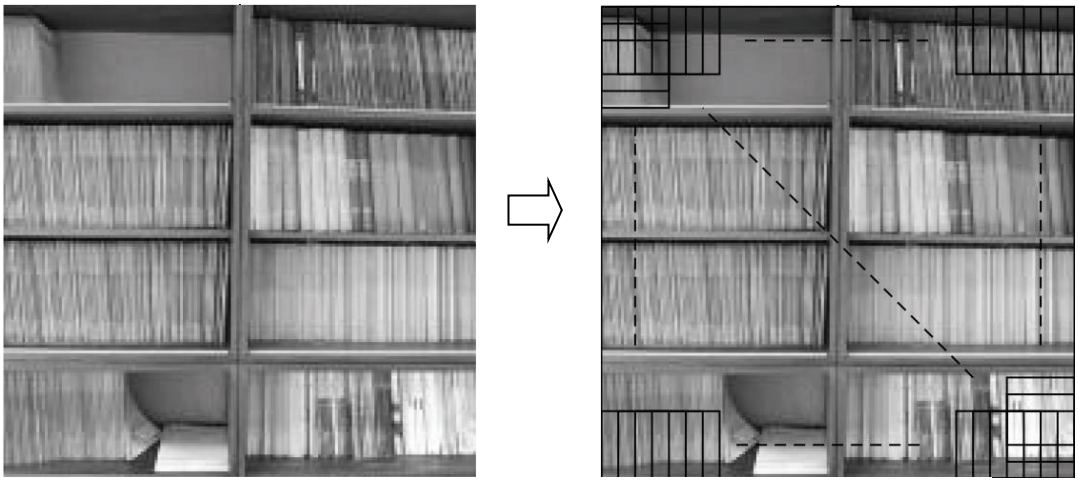
Hình 3.3. Kết quả thực nghiệm xử lý dữ liệu đầu vào

(a) Ảnh gốc *(b) Ảnh xám*

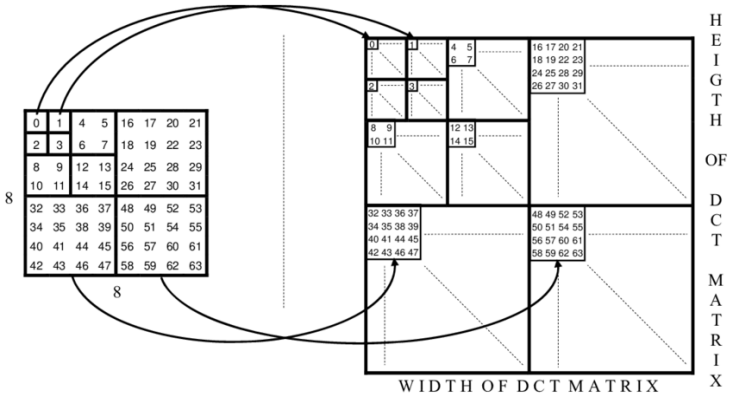
3.2.2. Xử lý tìm điểm cắt ghép trong từng khung hình

Quá trình xử lý tìm điểm cắt ghép trong từng khung hình của video được thực hiện theo một quy trình tuần tự. Cụ thể:

1) Chia ảnh xám đã được Module xử lý thực hiện thành các khối vuông điểm ảnh chồng nhau với kích thước cố định, ở đây ta áp dụng kích thước mỗi khối là 8x8. Các khối vuông điểm ảnh này được lấy từ góc trên màn hình bên trái theo thứ tự từ trên xuống dưới, từ trái qua phải đến góc dưới màn hình bên phải tạo thành các khối xếp chồng lên nhau. Mỗi khối được biểu diễn là điểm bắt đầu của hàng và cột tương ứng trong mỗi khung ảnh.



Hình 3.4 Chuyển từ ảnh xám sang các khối điểm ảnh 8x8



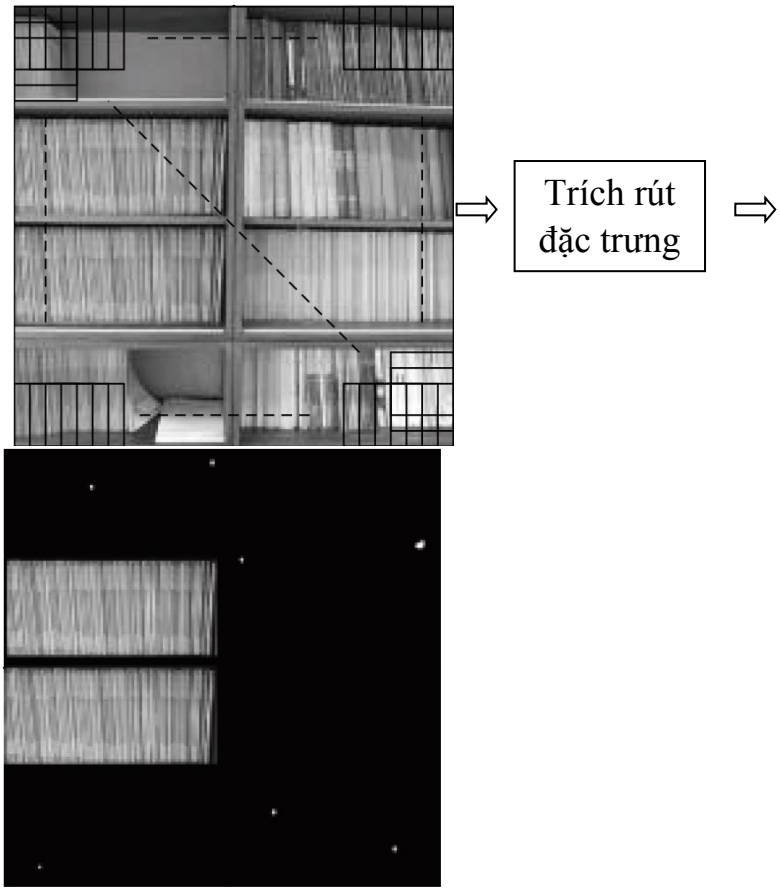
Hình 3.5. Chia các khung ảnh xám thành các khối kích thước 8x8 [8]

2) Áp dụng phép biến đổi Cosin rời rạc (DCT) cho mỗi khối điểm ảnh đã được trích xuất ra. Các khối điểm ảnh sẽ được gán vào trong một mảng 2 chiều, làm dữ liệu đầu vào cho hàm chức năng của phép biến đổi Cosin rời rạc.

3) Trích chọn đặc trưng từ các khối vuông đã áp dụng DCT. Thông qua phép biến đổi Cosin rời rạc, ta sẽ được một ma trận gồm các trọng số của DCT. Ma trận trọng số có thể được sắp xếp theo hình zig-zag để lấy các thông tin được lưu trong các khối đã hiển thị. Qua phép biến đổi Cosin rời rạc, mỗi khối sẽ được gán một trọng số nhất định làm cơ sở để tiến hành trích chọn các đặc trưng mỗi khối điểm ảnh.

0	1	5	6	14	15	27	28
2	4	7	13	16	26	29	42
3	8	12	17	25	30	41	43
9	11	18	24	31	40	44	53
10	19	23	32	39	45	52	54
20	22	33	38	46	51	55	60
21	34	37	47	50	56	59	61
35	36	48	49	57	58	62	63

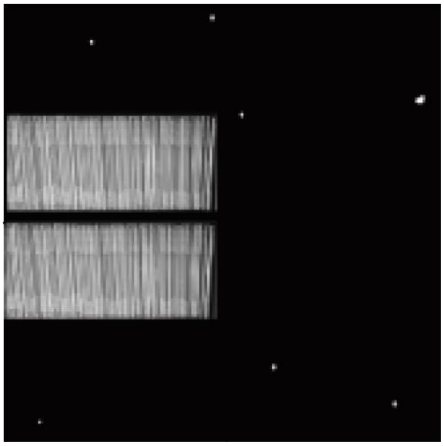
Hình 3.6. Các trọng số của ma trận DCT



Hình 3.7. Trích chọn đặc trưng, tìm kiếm và phát hiện các điểm trùng lặp

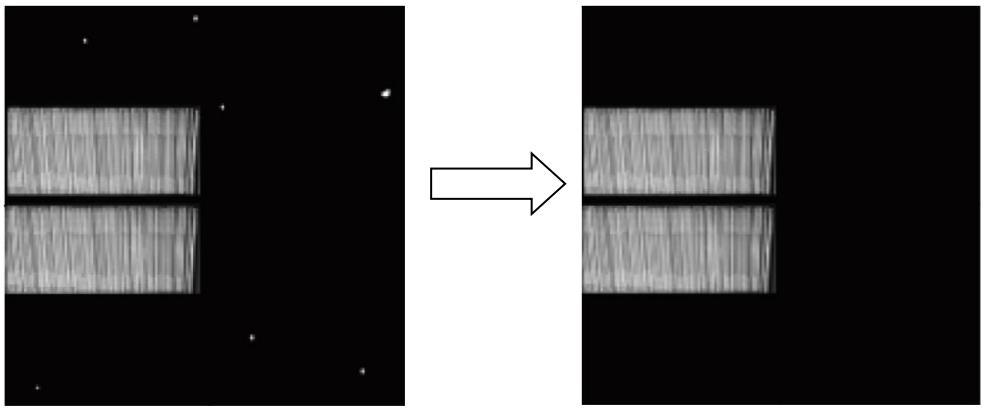
4) Dùng thuật toán học máy để ghép nối những điểm giống nhau theo tiêu chuẩn nhất định. Học viên lựa chọn các thuật toán phân cụm trong thư viện Scipy của Python để lọc những điểm có đặc trưng giống nhau thành các cụm. Cụ thể ở mô hình này, học viên đã sử dụng phương pháp phân tích thành phần chính PCA với hàm lõi Gaussian RBF để tối ưu hóa đặc trưng, giảm thiểu chiều dữ liệu, xác định số lượng nhóm nhất định. Qua nghiên cứu các thuật toán phân cụm dữ liệu trong học máy, học viên đã áp dụng phương pháp phân cụm K-nearest neighbor (viết tắt là KNN) vì tính đơn giản của nó. Kế thừa các nghiên cứu trong và ngoài nước về phân cụm dữ liệu trong xử lý ảnh, học viên đã linh hoạt áp dụng tính toán khoảng cách Eculid giữa các cụm dữ liệu đã được xử lý trước đó. Với một ngưỡng đã được nghiên cứu, tính toán cho trước, học viên đã áp dụng tính toán khoảng cách giữa các cụm dữ liệu, tiến hành phân loại thành các cụm có đặc điểm tương đồng, để đưa ra

những cụm dữ liệu có đặc điểm giống nhau, như một kết quả cho việc dự đoán các vùng bị chỉnh sửa trong mỗi khung ảnh của video.



Hình 3.8. Lọc những điểm có đặc trưng giống nhau thành các cụm

5) Cuối cùng ta sẽ tiến hành xóa bỏ những khối có kích thước nhỏ, nằm rời rạc nhau để đưa ra vùng dự đoán cuối cùng.



Hình 3.9. Xóa bỏ khối nhỏ, rời rạc

Đây là module quan trọng nhất trong chương trình phát hiện điểm cắt ghép trong video. Các tham số được sử dụng trong thuật toán gồm tham số lượng tử hóa, ngưỡng khoảng cách Eculid giữa các khối, ngưỡng khoảng cách Eculid giữa các pixel. Phụ thuộc từng tham số trên, hiệu suất tính toán cũng như kết quả đưa ra sẽ có sự khác nhau.

3.3. Kết quả thực nghiệm

Thuật toán được thử nghiệm trên 5 bộ video từ nguồn InVID Fake Video Corpus là những video đã được chỉnh sửa lấy trên nguồn các trang mạng xã hội

Youtube và Facebook có dung lượng từ 10-30s. Thiết bị thử nghiệm chạy thuật toán là máy tính có cấu hình Intel Core i7 4700MQ 2.4GHz, ổ cứng SSD 512GB, RAM 8GB để thử nghiệm khả năng tính toán của thuật toán. Ta sẽ thử nghiệm các bộ dữ liệu trên các giá trị khác nhau của phương pháp áp dụng bộ lọc Cosin rời rạc đó là:

- Kích thước các khối điểm ảnh: B
- Khoảng cách Eculid giữa các khối điểm ảnh: d_B
- Khoảng cách Eculid giữa các pixel: d_p
- Giới hạn vectơ dịch chuyển: Veclimit

Ta có các kết quả thử nghiệm như sau:

- Chuẩn bị dữ liệu:

Bảng 3.2. Bộ dữ liệu thực nghiệm

Bộ video	Dung lượng	fps	Số frame ảnh
Video 1	10s	23	234
Video 2	13s	24	312
Video 3	18s	22	400
Video 4	23s	23	531
Video 5	25	23	578

- Kết quả thử nghiệm B= 16x16, d_B=5, d_p=20, Vec_{limit}=20

Bảng 3.3. Kết quả thực nghiệm khối 16x16 pixels

Bộ video	Độ chính xác	TPR	FPR

Video 1	0.683	0.821	0.073
Video 2	0.667	0.805	0.082
Video 3	0.635	0.763	0.080
Video 4	0.598	0.725	0.092
Video 5	0.572	0.689	0.112

- Kết quả thử nghiệm B= 24x24, d_B=5, d_p=20, Vec_{limit}=20

Bảng 3.4. Kết quả thực nghiệm khối 24x24 pixels

Bộ video	Độ chính xác	TPR	FPR
Video 1	0.723	0.841	0.064
Video 2	0.687	0.725	0.092
Video 3	0.612	0.689	0.098
Video 4	0.598	0.693	0.106
Video 5	0.552	0.632	0.131

Qua hai lần thử nghiệm với hai bộ khối điểm ảnh khác nhau nhận thấy độ chính xác áp dụng mô hình trên có sự thay đổi. Ngoài sự ảnh hưởng việc thay đổi kích thước khối điểm ảnh trong thuật toán, độ chính xác còn phụ thuộc vào chất lượng video rất nhiều. Video càng có độ sắc nét cao thì khả năng chính xác cao hơn tuy nhiên thời gian thực hiện thuật toán lại khá lớn (trung bình 4 phút/khung ảnh). Tuy nhiên việc thực hiện thuật toán trên vẫn đảm bảo được việc phát hiện các video bị chỉnh sửa mặc dù độ chính xác vẫn chưa cao và thời gian thực hiện còn lớn.

3.4. Nhận xét

Qua nghiên cứu, thực nghiệm phương pháp trên, từ kết quả cho ta thấy được nhiều vấn đề cần phải giải quyết:

- *Về phương pháp,* đây là phương pháp đơn giản áp dụng việc xử lý từng khung ảnh trong video để tiến hành tìm ra vùng bị chỉnh sửa. Phương pháp sử dụng các biện pháp xử lý ảnh cơ bản và nâng cao, các thư viện sẵn có giúp cho việc lập trình dễ dàng hơn. Ngoài ra, phương pháp không đòi hỏi độ phức tạp trong xử lý dữ liệu, cũng như áp dụng các biện pháp học máy nâng cao để xây dựng module, do vậy giảm bớt được nhiều công đoạn tính toán phức tạp. Tuy nhiên, mặt hạn chế của phương pháp đó là thời gian xử lý mỗi khung ảnh khá dài, đặc biệt là với những video có chất lượng cao. Các vùng dự đoán đưa ra còn bị ảnh hưởng bởi nhiễu, phải chọn lọc các tham số đầu vào cho từn video để cho kết quả tốt nhất.
- *Về kết quả thử nghiệm,* bộ dữ liệu video thử nghiệm được lựa chọn là những video đã bị cắt ghép, chỉnh sửa. Qua quá trình thử nghiệm với các thông số khác nhau đã đánh giá được hiệu suất và khả năng của phương pháp được áp dụng. Độ chính xác của phương pháp giảm dần theo thời gian của mỗi video, video có thời gian càng ngắn thì độ chính xác càng cao và ngược lại. Đồng thời những video có độ phân giải cao sẽ có thời gian thực hiện lâu hơn, tuy nhiên kết quả lại chính xác hơn. Các tham số đầu vào của phương pháp cũng ảnh hưởng nhiều tới kết quả. Vì áp dụng việc chia các khối điểm ảnh do vậy với kích thước các khối điểm ảnh sẽ cho kết quả tính toán khác nhau, từ đó ảnh hưởng tới hiệu quả của phương pháp.

KẾT LUẬN

Sự xuất hiện của mạng xã hội nói chung và sự lan tỏa mạnh mẽ của các video đã làm thay đổi thế giới, hình thành một "thế giới ảo" đan xen với thế giới thực, tạo ra tương tác tối đa trong mọi quan hệ xã hội, vượt qua mọi khoảng cách về không gian và thời gian. Chính sự phát triển mạnh mẽ, đa dạng của các loại hình mạng xã hội tạo ra mạng lưới truyền thông đa phương tiện ngày càng hiện đại, để mọi người có thể trao đổi, tiếp cận thông tin nhanh chóng trên khắp thế giới, tạo điều kiện phát triển mọi mặt của đời sống xã hội; đồng thời, cũng tạo điều kiện thuận lợi để các thế lực phản động và tội phạm sử dụng vào các hoạt động vi phạm pháp luật, gây ra nguy cơ mất an ninh, an toàn thông tin, đặc biệt nguy hiểm khi chúng sử dụng để tuyên truyền, xuyên tạc thông tin, kích động biểu tình. Đáng chú ý, các đối tượng phạm tội đã và đang gia tăng các hoạt động giả mạo video để bịa đặt, xuyên tạc thông tin gây mất uy tín cá nhân, tổ chức, chính quyền; cũng như phục vụ các hành vi lừa đảo chiếm đoạt tài sản. Tại nước ta, công tác giám định video phục vụ xác thực tính chính xác của thông tin lan truyền trên internet và công tác điều tra, đấu tranh tội phạm còn nhiều hạn chế, việc ứng dụng các kỹ thuật, công nghệ hiện đại chưa đạt được nhiều thành tựu và phục vụ nhu cầu thực tiễn của xã hội. Do đó, yêu cầu cấp thiết đặt ra hiện nay là phải nghiên cứu, xây dựng các giải pháp để phát hiện điểm cắt, ghép trong video phục vụ công tác giám định hình sự.

Trong phạm vi nghiên cứu của luận văn, học viên đã trình bày một số vấn đề liên quan đến giám định video, trong đó tập trung nghiên cứu một số phương pháp tiếp cận giải quyết vấn đề bài toán đặt ra, từ đó đề xuất xây dựng chương trình thực nghiệm bằng biện pháp biến đổi Cosin rời rạc, thu được một số kết quả nhất định. Qua đó, học viên nhận thấy nghiên cứu, phát triển các giải pháp trong phát hiện điểm cắt ghép trong video là một hướng nghiên cứu mới, đáp ứng được tình hình thực tế trong thời kỳ khoa học công nghệ phát triển, các công cụ chỉnh sửa hình ảnh, video ngày càng phát triển, công nghệ chỉnh sửa tinh vi và hiện đại. Tuy nhiên, quá trình nghiên cứu đòi hỏi thời gian và thử nghiệm nhiều phương pháp khác nhau để

đưa ra hiệu quả tốt nhất. Bám sát mục tiêu, nhiệm vụ, sử dụng đúng đắn các phương pháp nghiên cứu khoa học, luận văn đã thu được một số thành công và về cơ bản đã giải quyết tốt mục tiêu, nhiệm vụ nghiên cứu đặt ra.

Trong tương lai, hướng nghiên cứu sâu hơn về việc phát hiện các điểm cắt ghép trong video nói riêng và phát hiện chỉnh sửa trong video nói chung cần xem xét đến sự thành công hiện có của các phương pháp tiến tiến hiện đại bắt nguồn từ thị giác máy tính, khai thác dữ liệu lớn và khoa học máy tính. Ví dụ các phương pháp về thống kê cho phép mô hình hóa hiệu quả thông tin quy mô lớn, do đó, đây có thể là một giải pháp giải quyết hiệu quả hiệu suất trong mô hình phát hiện giả mạo trong hình ảnh, video. Về việc giảm sự phức tạp trong việc trích xuất các đặc trưng trong video, vấn đề này có thể được điều chỉnh trong các bước lựa chọn các đặc trưng. Việc lưu ý trong việc phát hiện giả mạo trong video, việc thêm thông tin có thể cải thiện hiệu quả của mô hình. Thật vậy, các đối tượng cụ thể được phát hiện và được đặc tả trên cơ sở các đặc trưng (như màu sắc, hình dạng, kết cấu…), khi xem xét với các đặc trưng khác, hiệu quả của các đặc trưng này đem lại tốt hơn. Các đặc điểm khác không liên quan có thể là nhiễu, nó sẽ ảnh hưởng đến hiệu quả của mô hình. Giải quyết được những vấn đề này sẽ cải thiện rất nhiều hiệu suất cả về thời gian và các bước tính toán. Việc nghiên cứu, thử nghiệm các thuật toán để nâng cao hiệu quả trong bài toán phát hiện điểm cắt ghép trong video, hình ảnh, ứng dụng trong đời sống, đặc biệt là trong giám định chứng cứ pháp y, phục vụ cho lực lượng công an sẽ là hướng phát triển cần thiết và quan trọng sau này.

Luận văn là công trình nghiên cứu công phu, nghiêm túc, song do đây là vấn đề mới, rất khó và phức tạp, phạm vi nghiên cứu rộng, cộng thêm những khó khăn khách quan, kiến thức của học viên còn hạn chế nên chắc chắn còn nhiều khiếm khuyết. Học viên rất mong nhận được sự quan tâm, góp ý của các nhà khoa học, nhà hoạt động thực tiễn và đồng nghiệp. Cuối cùng, học viên xin chân thành cảm ơn các đơn vị liên quan, các đồng chí, đồng nghiệp, đặc biệt là thầy Phó Giáo sư Hà Hải Nam, người hướng dẫn khoa học đã tận tình giúp đỡ để học viên hoàn thành luận văn này./.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] Bestagini, P., Battaglia, S., Milani, S., Tagliasacchi, M., & Tubaro, S. (ICASSP 2013). *Detection of temporal interpolation in video sequences.* In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,.

[2] Bian, S., Luo, W., & Huang, J. (2014). *Exposing Fake Bit Rate Videos and Estimating Original Bit Rates.* IEEE Trans. Circuits Syst. Video Technol. .

[3] Chen, J. K. (2015). *Median filtering forensics based on convolutional neural networks.* IEEE Signal Processing Letters 22(11), 1849–1853.

[4] Chen, S., Tan, S., Li, B., & Huang, J. (2016). *Automatic Detection of Object-Based Forgery in Advanced Video.* IEEE Trans. Circuits Syst. Video Technol.

[5] Conotter, V., O'Brien, J., & Farid, H. (2012). *Exposing Digital Forgeries in Ballistic Motion.* . IEEE Trans. Inf. Forensics Secur. .

[6] Ding, X., Yang, G., Li, R., Zhang, L., Li, Y., & Sun, X. (2018). *Identification of Motion-Compensated Frame Rate Up-Conversion Based on Residual Signals.* . IEEE Trans. Circuits Syst. Video Technol. .

[7] Dirik, A. M. (2009). *Image tamper detection based on demosaicing artifacts.* In: Proc. of the 2009 IEEE International Conference on Image Processing (ICIP 2009), pp. 1497–1500.

[8] Fadl, S., Han, Q., & Li, Q. (2018). *Authentication of surveillance videos: Detecting frame duplication based on residual frame.* J. Forensic Sci.

[9] Farid, H. (2009). *Exposing digital forgeries from JPEG ghosts.* IEEE Transactions on Information Forensics and Security 4(1), 154–160.

[10] Ferrara, P. B. (2012). *Image forgery localization via fine-grained analysis of CFA artifacts.* IEEE Transactions on Information Forensics and Security 7(5).

[11] Gironi, A., Fontani, M., Bianchi, T., Piva, A., & Barni, M. (2014). *A video forensic technique for detecting frame deletion and insertion.* In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014.

[12] Gr´ egoire Mercier, F. M. (2019). Detecting manipulations in video.

[13] He, P., Jiang, X., Sun, T., & Wang, S. (2016). *Double compression detection based on local motion vector field analysis in static-background videos.* . J. Vis. Commun. Image Represent.

[14] He, P., Jiang, X., Sun, T., & Wang, S. (2017). *Detection of double compression in MPEG-4 videos based on block artifact measurement.* Neurocomputing .

[15] Hsu, C., Hung, T., Lin, C., & Hsu, C. (2008). *Video forgery detection using correlation of noise residue.* In Proceedings of the 2008 IEEE 10th Workshop on Multimedia Signal Processing, Cairns, QLD.

[16] J. Choi, B. T. (2014). *The placing task: A large-scale geo-estimation challenge for social-media videos and images.* In Proceedings of the 3rd ACM Multimedia Workshop on Geotagging and Its Applications in Multimedia.

[17] Johnson, J. (07/4/2021). *Worldwide digital population as of January 2021*.

[18] Kobayashi, M., Okabe, T., & Sato, Y. (2010). *Detecting Forgery From Static-Scene Video Based on Inconsistency in Noise Level Functions.* IEEE Trans. Inf. Forensics Secur.

[19] Liu, Y., & Huang, T. (2017). *Exposing video inter-frame forgery by Zernike opponent chromaticity moments and coarseness analysis.* Multimed. Syst.

[20] Mahdian, B. S. (2009). *Using noise inconsistencies for blind image forensics.* Image and Vision Computing 27(10), 1497–1503.

[21] Milani, S., Bestagini, P., Tagliasacchi, M., & Tubaro, S. (2012). *Multiple compression detection for video sequences.* In Proceedings of the 14th IEEE International Workshop on Multimedia Signal Processing, MMSP 2012.

[22] Nodegraph.se. (2020). *Retrieved from how much data is on the internet*. Nodegraph.se

[23] Qi, X. X. (2015). *A singular-value-based semi-fragile watermarking scheme for image content.* Journal of Visual Communication and Image Representation 30, 312–327.

[24] Qin, C. J. (2017). *Fragile image watermarking with pixel-wise.* Signal Processing 138, 280–293.

[25] Richao, C., Gaobo, Y., & Ningbo, Z. (2014). *Detection of object-based manipulation by the statistical features of object contour.* Forensic Sci. Int. .

[26] Sami Bourouis, R. A. (2020). *Recent Advances in Digital Multimedia Tampering Detection for Forensics Analysis.* Symmetry.

[27] Singh, G., & Singh, K. (2019). *Video frame and region duplication forgery detection based on correlation coefficient and coefficient of variation.* Multimed. Tools Appl.

[28] Soni, B. D. (2017). *CMFD: a detailed review of block based and key feature based techniques in image copy-move forgery detection.* IET Image Processing 12(2).

[29] Su, Y. X. (2010). *Detection of double compression in mpeg-2 videos.* International Workshop on Intelligent Syetems and Application (ISA).

[30] Thomas Mensink, R. B. (2017). *Spotting Audio-Visual Inconsistencies (SAVI) in Manipulated Video.* Open Access version.

[31] Ulutas, G., Ustubioglu, B., Ulutas, M., & Nabiyev, V. (2018). *Frame duplication detection based on BoW model.* Multimed. Syst. .

[32] Wang, W., & Farid, H. (2007). *Exposing Digital Forgeries in Interlaced and Deinterlaced Video.* . IEEE Trans. Inf. Forensics Secur.

[33] Warif, N. W. (2016). *Copy-move forgery detection: Survey, challenges and future directions.* Journal of Network and Computer Applications 100(75), 259–278.

[34] Yang, J., Huang, T., & Su, L. (2016). *Using similarity analysis to detect frame duplication forgery in videos.* Multimed. Tools Appl.

[35] Zhang, D., Yang, G., Li, F., Wang, J., & Sangaiah, A. (2020). *Detecting seam carved images using uniform local binary patterns.* . Multimed. Tools Appl.

[36] Zhang, Y. L. (2004). *Revealing the traces of median filtering using high-order local ternary patterns.* IEEE Signal Processing Letters 3(21), 275–279.

[37] Zhang, Z. H. (2015). Security and Communication networks 8(2).