

NGHIÊN CỨU HIỆU NĂNG CỦA CÁC PHƯƠNG PHÁP ĐÁNH GIÁ CHỦ QUAN VỚI DỮ LIỆU ẢNH VÀ VIDEO

Phí Công Huy

Khoa Đa ph ̣ ng ti ̣ n, H ̣ c Vi ̣ n Công Ngh ̣ B ̣ u Chính Vi ̣ n Thông

Abstract— Đối với lĩnh vực nghiên cứu về chất lượng ảnh video hiện nay, có nhiều phương pháp đánh giá chủ quan khác nhau, tùy thuộc vào mục đích. Tuy nhiên, theo khảo sát chưa có một nghiên cứu toàn diện về việc lựa chọn một phương pháp đánh giá chủ quan phù hợp và chính xác nhất. Do đó, bài báo này sẽ cung cấp một nghiên cứu toàn diện, dựa trên việc so sánh các phương pháp đánh giá chủ quan phổ biến nhất hiện nay. Trong bài báo này, tám phương pháp đánh giá chủ quan, bao gồm phương pháp đánh giá Thang đo liên tục (Double-Stimulus Continuous Quality-Scale (DSCQS)), phương pháp Thang đo suy giảm (Double-Stimulus Impairment Scale (DSIS)), phương pháp Xếp loại danh mục tuyệt đối với thang điểm 5 (ACR5), phương pháp Xếp loại danh mục tuyệt đối thang điểm 5 với tham chiếu âm (ACR5-HR), phương pháp Thang đo 11 mức độ ACR (ACR11), phương pháp Thang đo ACR11 với tham chiếu âm (ACR11-HR), phương pháp Đánh giá chủ quan chất lượng đa dạng của video (Subjective Assessment of Multimedia Video Quality (SAMVIQ)), và phương pháp SAMVIQ với tham chiếu âm (SAMVIQ-HR). Các phương pháp này được so sánh và kiểm thử theo từng tình huống, để xác định khả năng áp dụng của từng phương pháp đối với đánh giá ảnh và video. Các cách đánh giá bao gồm, đánh giá tổng thời gian và độ khó/dễ trong việc đánh giá đối với người tham gia, cũng như các đặc trưng của các thang đo khác nhau và độ tin cậy thống kê của từng phương pháp. Kết quả cho thấy hệ số tương quan và sự phân bố sự tương quan của điểm số bình quân (MOS) giữa các phương pháp là cao. Đồng thời, kết quả cũng thể hiện rõ các tham số độ tin cậy thống kê của từng phương pháp đánh giá như DSIS, ACR5, và SAMVIQ đã vượt trội hơn các phương pháp khác. Hơn nữa, kết quả khảo sát từ người tham gia cho thấy ACR5 là một phương pháp phù hợp nhất dựa trên tiêu chí về tổng thời gian đánh giá và độ dễ dàng trong việc đánh giá.

Keywords—Phương pháp đánh giá chủ quan, MOS, chất lượng ảnh và video, chất lượng trải nghiệm người dùng.

Tác giả liên hệ: Phí Công Huy,

Email: huyph@ptit.edu.vn

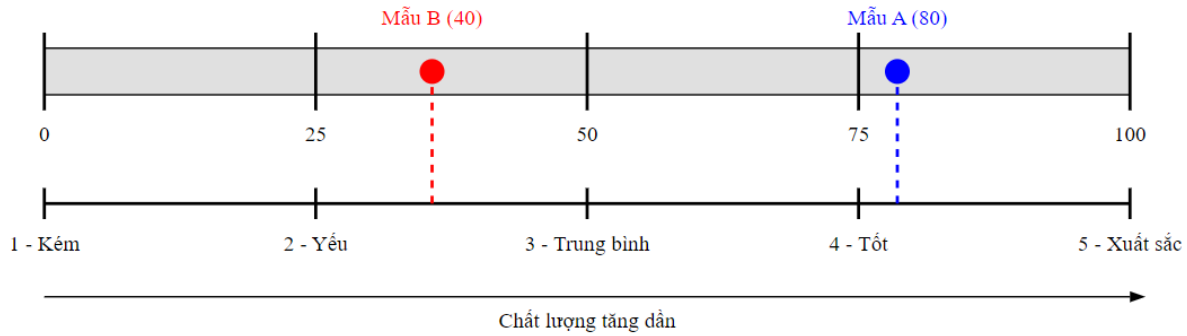
Đến tòa soạn: 8/2024, chỉnh sửa: 9/2024, chấp nhận đăng: 10/2024.

I. GIỚI THIỆU CHUNG

Dịch vụ Truyền hình Internet Protocol (IPTV) đang mở rộng, để cung cấp các dịch vụ này với chất lượng phù hợp, việc thiết kế các dịch vụ IPTV dựa trên chất lượng trải nghiệm của người dùng (QoE) rất quan trọng. Đánh giá chất lượng chủ quan là phương pháp cơ bản nhất để đánh giá QoE. Một số phương pháp đánh giá chủ quan phù hợp cho ứng dụng video đã được khuyến nghị bởi ITU-T và ITU-R. Phương pháp Double-Stimulus Continuous Quality-Scale (DSCQS) [1] là một phương pháp đánh giá chủ quan được sử dụng rộng rãi cho chất lượng video trong phát sóng và trực tuyến. Phương pháp Double-Stimulus Impairment Scale (DSIS) [1][2] (còn được gọi là phương pháp DCR hoặc EBU) và phương pháp Xếp loại Danh mục Tuyệt đối với thang điểm 5 (ACR5) [1][2] cũng được sử dụng. Ngoài ra, ACR với Tham chiếu Âm (ACR5-HR) [2] mới được phát triển để cân bằng tác động của sự đa dạng về chất lượng video gốc lên giá trị đánh giá chủ quan. Ngoài ra, phương pháp thang 11 điểm của ACR (ACR11) được liệt kê như một thang đo bổ sung có khả năng phân biệt cao hơn với phương pháp [2]. Hơn nữa, đánh giá chủ quan chất lượng đa dạng của video (SAMVIQ) [3] đã được phát triển như một phương pháp đánh giá chất lượng phù hợp cho nhiều định dạng video và môi trường đa phương tiện khác nhau, chẳng hạn như phát trực tuyến video cho máy tính cá nhân và thiết bị di động. Mặc dù tất cả các phương pháp này đều được tiêu chuẩn hóa trong ITU, nhưng sự khác biệt giữa các phương pháp như về độ nhạy và độ tin cậy chưa được nghiên cứu một cách đầy đủ. Một số nghiên cứu và so sánh hiệu năng của các phương pháp đã được thực hiện như: phương pháp DSCQS và phương pháp [4]; phương pháp ACR5-HR và DSCQS [5]; phương pháp SAMVIQ và DSCQS [6]; phương pháp ACR5 và SAMVIQ [7-8].

Tuy nhiên, đối với cùng một mục đích, các thang đo khác nhau sẽ được sử dụng trong các hoàn cảnh khác nhau, việc này làm cho việc so sánh chất lượng trong các thử nghiệm chủ quan khác nhau trở nên khó khăn, không chính xác.

Dựa vào thống kê trên, bài báo đưa ra một nghiên cứu tổng thể về các phương pháp đánh giá chủ quan, dựa trên việc xác định hiệu năng của tám phương pháp, gồm có phương pháp DSCQS, phương pháp DSIS, phương pháp ACR5, phương pháp ACR5-HR, phương pháp ACR11, phương pháp ACR11-HR, phương pháp SAMVIQ, và



Hình 2. Thang đo chất lượng của DSCQS

phương pháp SAMVIQ-HR. Các tiêu chí kiểm thử dựa trên thời gian đánh giá, sự khó khăn khi đánh giá, và độ chính xác của từng phương pháp. Đồng thời, dữ liệu đa dạng của video và ảnh sẽ được tổng hợp và đề xuất để đảm bảo sự khách quan của kết quả đánh giá, như: kho ảnh của Kaggle [9], kho ảnh Light Field [10], kho ảnh và video của VIS lab [11].

Cấu trúc bài báo gồm có 5 phần, với Phần 1 là giới thiệu chung về bài toán, các nghiên cứu liên quan về các phương pháp đánh giá chủ quan và nguồn dữ liệu được giới thiệu tại Phần 2. Phần 3 bài báo đề xuất các tiêu chí phân tích và kiểm thử các phương pháp đánh giá, sau đó là kết quả kiểm thử tại Phần 4. Cuối cùng, Phần 5 sẽ tổng kết lại toàn bộ nội dung của bài báo.

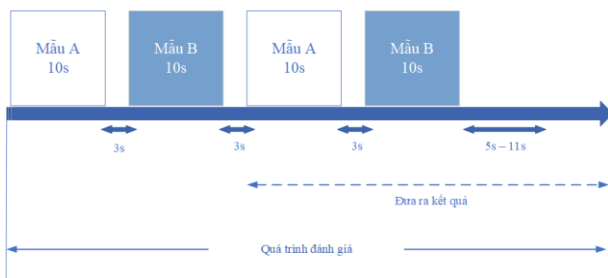
II. NHỮNG NGHIÊN CỨU LIÊN QUAN

Mục này mô tả tám phương pháp đánh giá chủ quan và điều kiện thử nghiệm đánh giá chủ quan được sử dụng trong thí nghiệm.

A. Các phương pháp đánh giá chủ quan

1. Phương pháp DSCQS

- Phương pháp đánh giá:** Phương pháp này yêu cầu đánh giá hai phiên bản của mỗi hình ảnh thử nghiệm. Một trong mỗi cặp hình ảnh thử nghiệm không bị suy giảm, trong khi hình ảnh còn lại có thể có hoặc không có sự suy giảm. Hình ảnh không bị suy giảm được sử dụng để làm tham chiếu, nhưng người xem không được thông báo. Trong quá trình thử nghiệm, vị trí của hình ảnh tham chiếu được thay đổi theo thứ tự ngẫu nhiên (Hình 1).



Hình 1. Phương pháp DSCQS

- Thước đo chất lượng:** Người xem chỉ cần đánh giá chất lượng tổng thể của mỗi cặp thử nghiệm bằng cách đánh dấu trên một thang đo. Thang đo được biểu diễn theo cặp để phù hợp với việc trình bày của mỗi bức

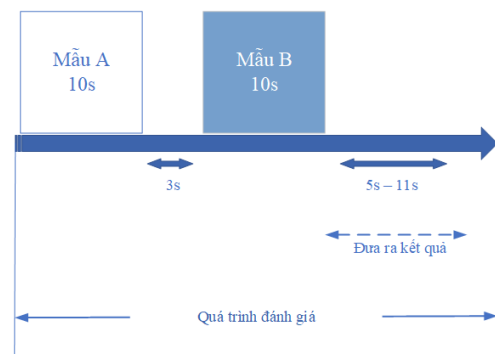
ảnh thử nghiệm. Các thang đo cung cấp một hệ thống xếp hạng liên tục để hạn chế việc ảnh hưởng sự thay đổi thị giác của người xem, các cặp được chia thành năm bậc có độ dài bằng nhau tương ứng với thang điểm được chia theo cấp chất lượng (xấu, kém, trung bình, tốt, xuất sắc). Các cặp đánh giá (tham chiếu và thử nghiệm) cho mỗi điều kiện thử nghiệm được chuyển đổi từ các cấp đo chất lượng thành các điểm số chuẩn hóa dao động từ 0 đến 100 (Hình 2).

- Cách tính DSCQS:** chất lượng được biểu thị bằng điểm số của người xem khác nhau (D) phân biệt dựa trên điểm số của người xem (V) đối với chuỗi video đã qua xử lý (Processed Video - PV) và tham chiếu (REF)

$$D(PV) = |V(PV) - V(REF)| \quad (1)$$

2. Phương pháp DSIS

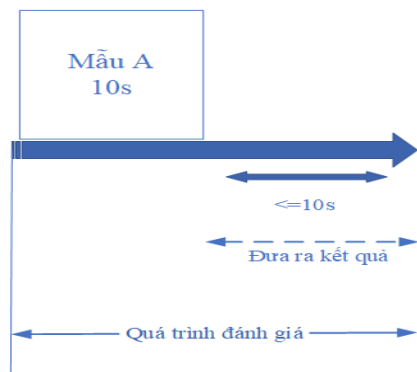
- Phương pháp đánh giá:** Phương pháp DSIS mô tả các chuỗi thử nghiệm được trình bày theo cặp: mẫu đánh giá đầu tiên được trình bày trong mỗi cặp luôn là tham chiếu gốc, trong khi mẫu đánh giá thứ hai là các mẫu video/ảnh cùng một nguồn được trình bày thông qua công cụ biểu diễn (Hình 3).



Hình 3. Phương pháp DSIS

- Thang đo chất lượng:** Mỗi người xem được chọn một loại trong năm loại cảm xúc (1: rất gây khó chịu, 2: gây khó chịu, 3: hơi gây khó chịu, 4: nhận biết được nhưng không gây khó chịu, 5: không nhận biết được).

- **Cách tính DSIS:** Chất lượng được biểu thị bằng điểm số của người xem với đoạn video/ảnh đã được xử lý $V(PV)$.
- 3. **Phương pháp ACR5/ACR5-HR**
- **Phương pháp đánh giá:** Phương pháp ACR là một dạng phán đoán theo danh mục, trong đó các chuỗi kiểm tra được trình bày lần lượt (Hình 4).

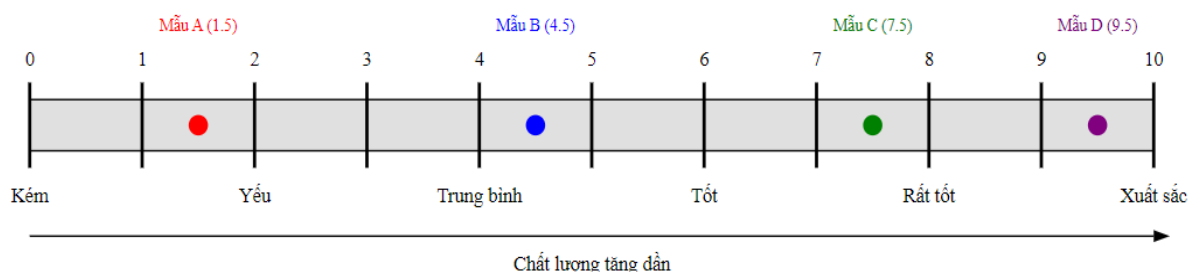


Hình 4. Phương pháp ACR5 và ACR11

- **Thang đo chất lượng:** Mỗi người xem chọn một danh mục từ thang điểm rời rạc gồm năm danh mục (1: xấu, 2: kém, 3: trung bình, 4: tốt, 5: tuyệt vời).
- **Cách tính ACR5:** Chất lượng được biểu thị bằng điểm số của người xem với đoạn video/ảnh đã được xử lý $V(PV)$.
- **Cách tính ACR5-HR:** Chất lượng được biểu diễn bằng điểm của người xem khác nhau (D) như sau, sử dụng chuỗi video đã xử lý (PV) và tham chiếu ảnh ($HREF$).

$$D(PV) = V(PV) - V(HREF) + 5 \quad (2)$$

- 4. **Phương pháp ACR11/ACR11-HR**
- **Phương pháp đánh giá:** Phương pháp ACR11 là một đánh giá theo hạng mục, trong đó các chuỗi kiểm tra được trình bày từng mẫu.
- **Thang đo chất lượng:** Mỗi người xem chọn một điểm từ thang điểm rời rạc gồm 11 điểm trong 5 hạng mục (Hình 5).



Ghi chú: Thang điểm 0-10, đánh giá chi tiết đến 0.5 điểm | A (1.5): Gần kém | B (4.5): Trung bình khá | C (7.5): Tốt | D (9.5): Gần hoàn hảo

Ứng dụng: Đánh giá chất lượng âm thanh, video, trải nghiệm người dùng, và các lĩnh vực khác cần đánh giá chi tiết.

Hình 5. Thang điểm của ACR11

- **Cách tính ACR11:** Chất lượng được biểu thị bằng điểm số của người xem với đoạn video/ảnh đã được xử lý $V(PV)$.
- **Cách tính ACR11-HR:** Chất lượng được biểu diễn bằng điểm của người xem khác nhau (D) như sau, sử dụng chuỗi video đã xử lý (PV) và tham chiếu ảnh ($HREF$).

$$D(PV) = V(PV) - V(HREF) + 10 \quad (3)$$

5. Phương pháp SAMVIQ/SAMVIQ-HR

- **Phương pháp đánh giá:** Trong phương pháp này, người xem được truy cập vào nhiều phiên bản của một chuỗi video kiểm thử. Khi tất cả các mẫu đã được người xem đánh giá, nội dung của chuỗi tiếp theo mới được biểu diễn. Các phiên bản khác nhau được người xem lựa chọn ngẫu nhiên thông qua công cụ biểu diễn. Người xem có thể dừng, xem lại, và chỉnh sửa điểm số của mỗi phiên bản theo ý muốn. Việc đánh giá chất lượng được thực hiện từng cảnh (xem Hình 6), bao gồm một tham chiếu gốc và một tham chiếu ẩn.



Hình 6. Phương pháp thể hiện SAMVIQ

- **Thang đo chất lượng:** Phương pháp này sử dụng thang đo chất lượng liên tục để cung cấp một kết quả về chất lượng tổng quan của các chuỗi video. Mỗi người xem di chuyển thanh trượt trên thang đo liên tục từ 0 đến 100 được chú thích bởi 5 mục chất lượng sắp

xếp theo thứ tự tuyến tính (xấu, kém, trung bình, tốt, xuất sắc).

- **Cách tính SAMVIQ:** Chất lượng được biểu thị bằng điểm số của người xem với đoạn video/ảnh đã được xử lý $V(PV)$.
- **Cách tính SAMVIQ-HR:** Chất lượng được biểu diễn bằng điểm của người xem khác nhau (D) như sau, sử dụng chuỗi video đã xử lý (PV) và tham chiếu ảnh ($HREF$).

$$D(PV) = V(PV) - V(HREF) + 100 \quad (4)$$

B. Các thể loại dữ liệu

Để phục vụ các kiểm thử, dữ liệu đóng vai trò vô cùng quan trọng, bởi nhóm dữ liệu cần phải đủ rộng, linh hoạt, các tiêu chí của dữ liệu cần đảm bảo tính khách quan nhất. Trong bài báo này, 3 nhóm dữ liệu được sử dụng khác nhau và lấy từ các tổ chức uy tín, cụ thể là nhóm dữ liệu từ Kaggle, nhóm dữ liệu Light Field, và nhóm dữ liệu từ phòng nghiên cứu VIS lab.

1. Dữ liệu từ Kaggle

Kaggle là một nền tảng trực tuyến hàng đầu trong lĩnh vực khoa học dữ liệu và học máy, cung cấp một môi trường nghiên cứu cho các nhà khoa học dữ liệu, kỹ sư học máy và các chuyên gia trong lĩnh vực này. Trong [12] đề xuất một phương pháp đào tạo kỹ năng khoa học dữ liệu dựa trên các cuộc thi có sử dụng bộ dữ liệu của Kaggle. Bài báo [13] nhấn mạnh vai trò của Kaggle trong việc tạo ra các cuộc thi học máy học thuật, nơi một phần dữ liệu được giữ lại để kiểm tra. Phương pháp này đảm bảo việc xếp hạng khách quan các thí sinh và tạo ra một môi trường học tập trải nghiệm trong khoa học dữ liệu. Dữ liệu trên Kaggle rất linh hoạt, và điều này thể hiện qua một số khía cạnh chính:

- Đa dạng về chủ đề: Kaggle cung cấp dữ liệu từ nhiều lĩnh vực khác nhau như y tế, tài chính, thương mại điện tử, khoa học xã hội, và nhiều lĩnh vực khác. Điều này cho phép các nhà khoa học dữ liệu và kỹ sư học máy thử nghiệm và phát triển mô hình trong nhiều bối cảnh khác nhau.
- Định dạng dữ liệu phong phú: Các bộ dữ liệu trên Kaggle có thể ở nhiều định dạng khác nhau, bao gồm CSV, JSON, hình ảnh, văn bản, và âm thanh. Điều này giúp người dùng có thể làm việc với các loại dữ liệu khác nhau mà không cần phải chuyển đổi định dạng.
- Quy mô dữ liệu: Kaggle có các bộ dữ liệu với quy mô từ nhỏ đến rất lớn, giúp người dùng có thể lựa chọn dữ liệu phù hợp với khả năng tính toán và nhu cầu phân tích của mình. Điều này đặc biệt hữu ích cho việc thử nghiệm các mô hình học máy trên các tập dữ liệu có kích thước khác nhau.
- Cộng đồng hỗ trợ: Kaggle có một cộng đồng lớn các nhà khoa học dữ liệu sẵn sàng chia sẻ mã nguồn, phương pháp và kết quả. Điều này giúp người dùng học hỏi từ những người khác và nhanh chóng áp dụng các kỹ thuật mới vào các bộ dữ liệu khác nhau.
- Tính khả dụng và truy cập dễ dàng: Các bộ dữ liệu trên Kaggle thường miễn phí và có thể được truy

cập dễ dàng, điều này tạo điều kiện thuận lợi cho việc học tập và nghiên cứu.

Những yếu tố trên làm cho dữ liệu trên Kaggle trở thành một công cụ hữu ích cho cả người mới bắt đầu và các chuyên gia trong lĩnh vực khoa học dữ liệu và học máy.

2. Dữ liệu Light Field

Light Field là một khái niệm quan trọng trong lĩnh vực thị giác máy tính và xử lý ảnh, cung cấp thông tin phong phú về cấu trúc 3D của cảnh và các đặc tính quang học. Light Field là một cách biểu diễn đầy đủ thông tin về ánh sáng trong không gian, bao gồm cả cường độ và hướng của các tia sáng. Khái niệm này được giới thiệu lần đầu bởi Levoy và Hanrahan vào năm 1996, mở ra một hướng mới trong việc thu thập và xử lý thông tin hình ảnh. Dữ liệu Light Field cung cấp một cách tiếp cận mới để hiểu và phân tích cảnh 3D, vượt qua những hạn chế của phương pháp chụp ảnh truyền thống. Dữ liệu Light Field đại diện cho một bước tiến quan trọng trong lĩnh vực thị giác máy tính và xử lý ảnh. Với khả năng cung cấp thông tin phong phú về cấu trúc 3D và đặc tính quang học của cảnh, dữ liệu Light Field mở ra nhiều cơ hội mới trong các ứng dụng như tái tạo 3D, tái lấy nét sau chụp, và ước tính độ sâu.

Phương pháp thu thập: Dữ liệu light field có thể được thu thập bằng nhiều phương pháp khác nhau, mỗi phương pháp tạo ra dữ liệu với đặc điểm riêng:

- Mảng camera: Cung cấp dữ liệu từ nhiều góc nhìn rời rạc, thường có độ phân giải cao. b) Camera plenoptic: Tạo ra dữ liệu với thông tin hướng ánh sáng chi tiết trong một lần chụp. c) Quét với camera đơn: Tạo ra dữ liệu từ nhiều vị trí khác nhau theo thời gian. d) Tổng hợp máy tính: Tạo ra dữ liệu light field từ mô hình 3D, cho phép kiểm soát hoàn toàn các tham số.
- Cấu trúc dữ liệu: Dữ liệu light field có thể được biểu diễn dưới nhiều dạng cấu trúc: a) 4D Light Field (4D LF): Biểu diễn phổ biến nhất, sử dụng hai mặt phẳng song song. b) Lumigraph: Biến thể của 4D LF, thêm thông tin về hình dạng đối tượng. c) Surface Light Field: Biểu diễn ánh sáng phát ra từ bề mặt đối tượng. d) Unstructured Light Field: Dữ liệu từ các vị trí camera không theo quy tắc.
- Độ phân giải và mật độ: Dữ liệu light field có thể có độ phân giải và mật độ góc nhìn khác nhau: a) Độ phân giải không gian: Từ thấp (vài trăm pixel) đến cao (hàng chục megapixel). b) Mật độ góc: Từ thưa (vài góc nhìn) đến dày đặc (hàng trăm góc nhìn). c) Độ phân giải góc: Liên quan đến khả năng phân biệt các hướng ánh sáng.
- Phạm vi cảnh: Dữ liệu light field có thể bao gồm nhiều loại cảnh khác nhau: a) Cảnh trong nhà: Văn phòng, phòng khách, v.v. b) Cảnh ngoài trời: Phong cảnh, đô thị, v.v. c) Đối tượng đơn lẻ: Sản phẩm, tác phẩm nghệ thuật, v.v. d) Cảnh động: Bao gồm chuyển động của đối tượng hoặc camera.

- Ứng dụng chuyên biệt: Dữ liệu light field được thu thập và sử dụng cho nhiều ứng dụng khác nhau: a) Y tế: Light field nội soi, chụp cắt lớp quang học. b) Công nghiệp: Kiểm tra chất lượng, đo lường 3D. c) Giải trí: Điện ảnh, thực tế ảo và thực tế tăng cường. d) Khoa học: Hiển vi light field, thiên văn học.

Sự đa dạng này của dữ liệu light field không chỉ thể hiện qua phương pháp thu thập và cấu trúc dữ liệu, mà còn qua nội dung, chất lượng và mục đích sử dụng. Điều này tạo ra một lĩnh vực nghiên cứu phong phú với nhiều thách thức và cơ hội trong việc phát triển các thuật toán và ứng dụng mới.

3. Dữ liệu VIS lab

Video and Image Sense Lab (VIS Lab) thường tập trung vào nghiên cứu và phát triển các công nghệ liên quan đến xử lý hình ảnh và video. Dữ liệu của VIS Lab có thể bao gồm:

- Hình ảnh tĩnh:
 - Phân loại hình ảnh: Dữ liệu hình ảnh được sử dụng để huấn luyện các mô hình phân loại, giúp máy tính nhận diện và phân loại các đối tượng trong ảnh.
 - Phân đoạn ảnh: Dữ liệu này giúp chia nhỏ hình ảnh thành các phần khác nhau, thường để xác định ranh giới của các đối tượng cụ thể.

- Nhận diện khuôn mặt: Sử dụng dữ liệu hình ảnh khuôn mặt để phát triển các hệ thống nhận diện và xác thực danh tính.

▪ Video:

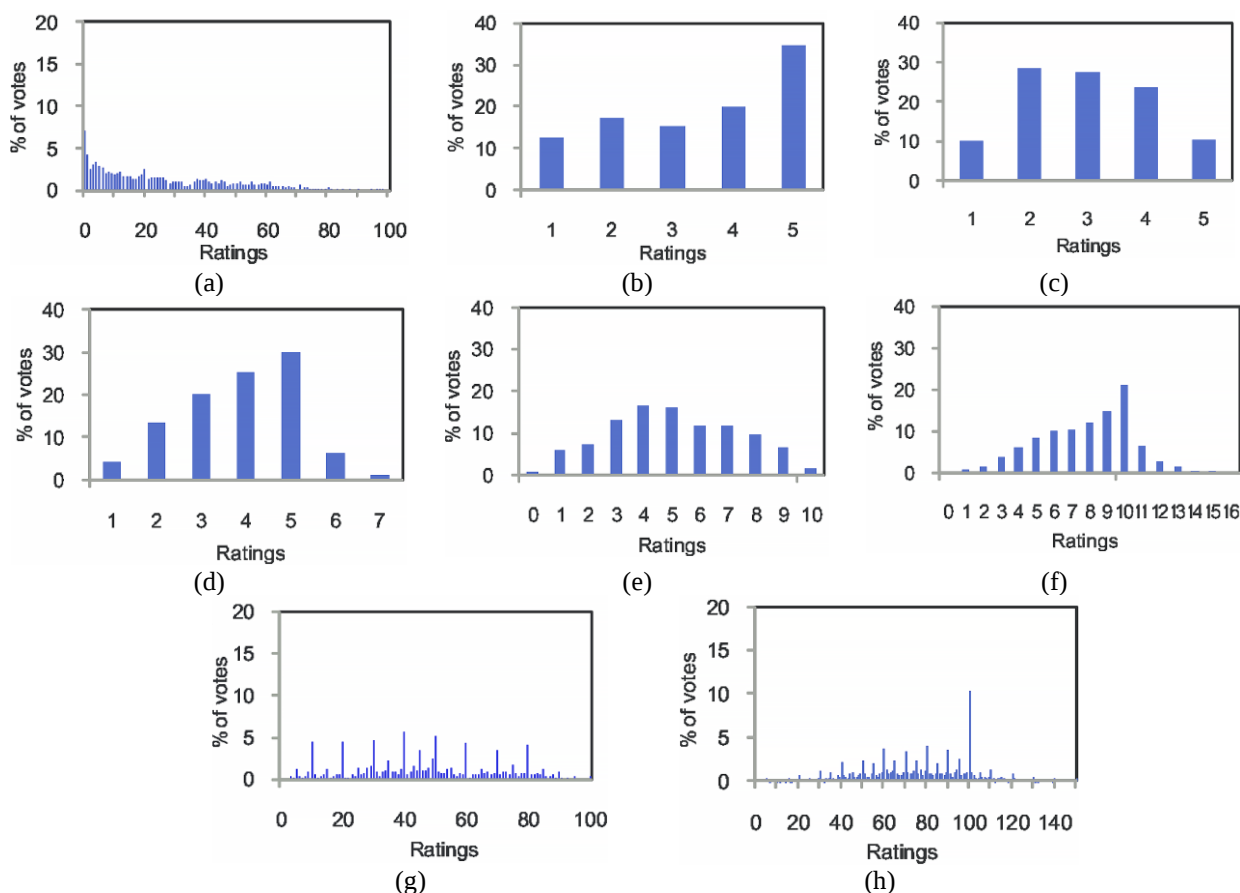
- Theo dõi đối tượng: Các bộ dữ liệu video giúp phát triển các thuật toán theo dõi đối tượng qua nhiều khung hình.
- Phân tích hành vi: Dữ liệu video được sử dụng để nhận diện và phân tích các hành động hoặc sự kiện diễn ra trong video.
- Nhận diện hoạt động: Phát triển các mô hình để nhận diện và phân loại các hoạt động của con người trong video.

▪ Dữ liệu gắn nhãn:

- Các bộ dữ liệu này thường bao gồm hình ảnh hoặc video đã được gắn nhãn thủ công, cung cấp thông tin chính xác về nội dung của từng khung hình hoặc hình ảnh, chẳng hạn như loại đối tượng, vị trí, và hành động.

▪ Dữ liệu 3D:

- Đám mây điểm (Point Cloud): Dữ liệu này thường được thu thập từ các cảm biến như LiDAR, giúp nghiên cứu các ứng dụng như tự động lái xe và mô phỏng môi trường 3D.



Hình 7. Phân bố thang điểm dựa trên kết quả đánh giá của từng phương pháp, theo thứ tự: a) DSCQS, b) DSIS, c) ACR5, d) ACR5-HR, e) ACR11, f) ACR11-HR, g) SAMVIQ, h) SAMVIQ-HR

- Mô hình 3D: Sử dụng trong các ứng dụng như thực tế ảo hoặc tăng cường, nơi cần tái tạo lại các đối tượng hoặc môi trường một cách chính xác.
- Dữ liệu từ cảm biến phụ trợ:
 - Kết hợp dữ liệu từ các cảm biến khác như âm thanh, nhiệt độ, hoặc chuyển động để cung cấp thêm thông tin ngữ cảnh cho các phân tích đa phương tiện.
- Dữ liệu từ người dùng:
 - Dữ liệu này có thể bao gồm tương tác của người dùng với các hệ thống thị giác máy tính, được sử dụng để cải thiện giao diện người dùng và tối ưu hóa hiệu suất của các thuật toán.

Các phòng thí nghiệm như VIS Lab thường sử dụng các bộ dữ liệu công khai nổi tiếng như ImageNet, COCO, hoặc KITTI, và cũng có thể tự phát triển các bộ dữ liệu riêng cho các nghiên cứu đặc thù.

III. ĐỀ XUẤT PHƯƠNG PHÁP VÀ HÌNH THỨC ĐÁNH GIÁ CÁC PHƯƠNG PHÁP

A. Điều kiện kiểm thử

Thử nghiệm đánh giá chất lượng chủ quan được thực hiện cho các loại video và hình ảnh dựa trên dataset tại Mục B Phần 2. Các nguồn video/ảnh được chuyển đổi và xử lý kích thước đồng bộ thành QVGA (320 x 240 pixel) và mã hóa bằng chuẩn mã hóa video H.266. Trong bài báo đã lựa chọn sử dụng 42 PV với đa dạng thể loại, bao gồm 5 video nguồn và 50 video đã mã hóa bao gồm các video đã xử lý. Từ các video nguồn sẽ được mã hóa để đảm bảo có chất lượng khác nhau về độ méo, tiếp theo, các video đã được xử lý sẽ sắp xếp ngẫu nhiên để lựa chọn ra 42 PV phù hợp thực hiện đánh giá. Các yếu tố xử lý trong video gồm giảm chất lượng tốc độ bit, tốc độ khung hình, nhiều bộ song, và độ dài từng khung hình tĩnh. Chi tiết về điều kiện chất lượng video được liệt kê trong Bảng 1. Video được biểu diễn với màn hình 24-inch độ phân giải FHD (1280 x 1024 pixel). Các điều kiện hiện thị được liệt kê trong Bảng 2.

Bảng 1. Các thông số về chất lượng video

Số lượng video thực hiện (Kaggle, Light Field, VIS lab)	Kaggle: 10 video Light Field: 20 video VIS lab: 10 video
Độ dài	15 giây
Mã hóa	H.266/VVC
Định dạng	QVGA (320 x 240)
Bitrate	22, 32, 40, 44
Frame rate	30 fps
Nhiều bộ song	Nhiều hạt tiêu
Khung hình tĩnh	Số lượng khung hình tĩnh 1 khung hình, cho các frame 1, 5, 15.

Bảng 2. Các điều kiện hỗ trợ người xem

Thiết bị	Màn hình 27" của Dell
Độ phân giải	Full HD (1920 x 1080)
Độ sáng	400 Nits cd/m ²
Tần số quét	60 Hz
Khoảng cách nhìn tới màn	50 cm
Môi trường	Phòng kín, ánh sáng không phản chiếu

Bảng 3. Hệ số tương quan r giữa các phương pháp

	DSCQS	DSIS	ACR5	ACR5-HR	ACR11	ACR11-HR	SAMVIQ	SAMVIQ-HR
DSCQS	-	-0.98	-0.98	-0.98	-0.98	-0.98	-0.98	-0.97
DSIS	-0.98	-	0.98	0.96	0.98	0.97	0.98	0.97
ACR5	-0.98	0.98	-	0.97	0.98	0.97	0.98	0.97
ACR5-HR	-0.98	0.96	0.97	-	0.97	0.98	0.97	0.98
ACR11	-0.98	0.98	0.97	0.97	-	1.00	0.98	0.97
ACR11-HR	-0.98	0.97	0.98	0.98	1.00	-	0.97	0.97
SAMVIQ	-0.98	0.98	0.98	0.97	0.98	0.97	-	0.99
SAMVIQ-HR	-0.97	0.97	0.97	0.98	0.97	0.97	0.99	-
TRUNG BÌNH	-0.98	0.98	0.98	0.97	0.98	0.98	0.98	0.98

Bảng 4. Hệ số tương quan spearman giữa các phương pháp

	DSCQS	DSIS	ACR5	ACR5-HR	ACR11	ACR11-HR	SAMVIQ	SAMVIQ-HR
DSCQS	-	0.98	0.97	0.98	0.97	0.97	0.98	0.97
DSIS	0.98	-	0.97	0.97	0.97	0.97	0.98	0.97
ACR5	0.97	0.97	-	0.97	0.98	0.97	0.99	0.97
ACR5-HR	0.98	0.97	0.97	-	0.99	0.98	0.96	0.98
ACR11	0.97	0.97	0.98	0.99	-	0.99	0.97	0.96
ACR11-HR	0.97	0.97	0.97	0.98	0.99	-	0.96	0.97
SAMVIQ	0.98	0.98	0.99	0.96	0.97	0.96	-	0.98
SAMVIQ-HR	0.97	0.97	0.97	0.98	0.96	0.97	0.98	-
TRUNG BÌNH	0.98	0.97	0.97	0.97	0.98	0.97	0.98	0.97

Bảng 5. mối tương quan giữa MOS_{range} và MCI_{norm} cho từng phương pháp

	DSCQS	DSIS	ACR5	ACR5-HR	ACR11	ACR11-HR	SAMVIQ	SAMVIQ-HR
MCI	4.58	0.25	0.23	0.28	0.47	0.56	4.26	4.87
MOS_{range}	53.54	3.56	3.04	3.08	5.77	5.56	59.21	58.13
MCI_{norm}	0.09	0.07	0.07	0.09	0.08	0.10	0.07	0.08

Bảng 6. thời gian đánh giá (phút) cho từng phương pháp

	DSCQS	DSIS	ACR5	ACR5-HR	ACR11	ACR11-HR	SAMVIQ	SAMVIQ-HR
MAX	45	25	18	18	20	20	38	38
MIN	37	16	11	11	11	11	23	23
AVERAGE	41	20	12	12	14	14	29	29
STD.	2	2	1	1	3	3	5	5

Người tham gia kiểm thử gồm có 40 người từ 19 đến 39 tuổi. Người tham gia được chia thành 2 nhóm lớn, mỗi nhóm 20 người (Nhóm A và Nhóm B). Sau đó, mỗi nhóm này tiếp tục được chia thành 5 nhóm nhỏ với số lượng 4 người/nhóm. Trong khoảng thời gian đánh giá, theo từng buổi (sáng và chiều), một nhóm từ Nhóm A và một nhóm từ Nhóm B được kết hợp tham gia cùng nhau, tương tự như vậy các nhóm còn lại thay đổi cho tới cuối quá trình đánh giá.

Năm phương pháp được thay đổi thứ tự ngẫu nhiên, đảm bảo không có sự trùng lặp. Cả hai nhóm sử dụng cùng một phương pháp trong ngày nhưng xem các video xử lý ngẫu nhiên. Thời gian thí nghiệm của mỗi phương pháp được đo và ghi điểm theo đánh giá của người tham gia sau khi kết thúc mỗi buổi thử nghiệm.

B. Phân tích các yếu tố liên quan

Phần này mô tả kết quả so sánh hiệu suất dựa trên các giá trị đánh giá chủ quan. Ba chỉ số hiệu suất, hệ số tương quan, hệ số tương quan thứ tự, và khoảng tin cậy trung bình, đã được sử dụng.

1. Mối quan hệ giữa điểm trung bình của DSCQS và các phương pháp khác

Phân phối điểm số cho từng phương pháp được thể hiện ở Hình 7. Trong thử nghiệm này, có nhiều video đã xử lý có chất lượng video gần như giống với video tham chiếu (tức là có điểm chất lượng video gần như giống nhau). Do đó, tỷ lệ phần trăm của các điểm cao nhất được gán theo thứ tự sau 5, 5, 10 và 100 tương ứng cho DSIS, ACR5-HR, ACR11-HR và SAMVIQ-HR. Dựa trên các kết quả này, bài báo đã so sánh kết quả dựa trên điểm số ý kiến trung bình (MOS) cho 42 PV. Các mối quan hệ giữa DSCQS và các phương pháp khác được thể hiện dựa trên tính toán các hệ số tương quan (R) và hệ số tương quan xếp hạng của Spearman, để làm rõ các mối quan hệ và đặc điểm của từng loại đánh giá. Bảng 3 và 4 liệt kê R và Spearman giữa các phương pháp. Tất cả các R cho DSCQS đều âm. Điều này thể hiện DSCQS cho giá trị nhỏ hơn khi chất lượng cao hơn, trái ngược với bảy phương pháp khác, giá trị nhỏ hơn khi chất lượng thấp hơn. Giá trị trung bình tuyệt đối của R và Spearman cho từng phương pháp được liệt kê ở dòng cuối cùng của Bảng 3 và 4. Không có sự khác biệt đáng kể với mức ý nghĩa 5% trong các giá trị trung bình của R và Spearman cho từng phương pháp. Điều này có nghĩa là mỗi

phương pháp cho thấy khả năng đánh giá giống nhau trong R và Spearman.

2. Phân tích độ tin cậy thống kê của điểm trung bình

Để làm rõ sự ổn định và sự khác biệt của việc đánh giá chất lượng, giá trị trung bình của khoảng tin cậy 95% (MCI) cho 42 MOS đã được so sánh. Với quy định MCI càng nhỏ, thì độ ổn định và sự khác biệt càng cao. Mỗi phương pháp có thang đo chất lượng khác nhau. Do đó, cần phải chuẩn hóa MCI để so sánh từng phương pháp. Bảng 5 liệt kê MCI và phạm vi MOS cho từng phương pháp. Ví dụ, các phương pháp DSCQS và SAMVIQ sử dụng cùng thang đo chất lượng 100 cấp, nhưng phạm vi điểm trung bình khác nhau. Vì vậy, kết quả cần được chuẩn hóa MCI bằng cách sử dụng phạm vi MOS cho các thang đo khác nhau. Cách tính của MCI_{norm} như sau:

$$MCI_{norm} = MCI / MOS_{range} \quad (5)$$

Bảng 5 cũng liệt kê MCI_{norm} cho từng phương pháp. DSIS, ACR5, và SAMVIQ có MCI_{norm} nhỏ nhất trong số 8 phương pháp. MCI_{norm} của ACR5-HR, ACR11-HR, và SAMVIQ-HR lớn hơn so với ACR5, ACR11, và SAMVIQ, tương ứng.

IV. KẾT QUẢ VÀ KIỂM THỬ

Phần này mô tả kết quả so sánh thời gian đánh giá tổng cộng và độ dễ dàng đánh giá từ quan điểm hiệu quả thí nghiệm và ý kiến của người xem.

A. Tổng thời gian thử nghiệm

Đánh giá chất lượng chủ quan đòi hỏi nhiều thời gian và chi phí. Thời gian đánh giá ngắn hơn là mong muốn nếu độ chính xác đánh giá vẫn như cũ. Đánh giá chất lượng chủ quan được thực hiện bởi mỗi nhóm con gồm bốn người xem cùng một lúc. Về tổng thời gian đánh giá, thời gian trình chiếu video là cố định nhưng thời gian bình chọn bị giới hạn tối đa 11 giây. Trong trường hợp các phương pháp DSCQS, DSIS, ACR5, và ACR11, nếu các nhóm nhỏ hoàn thành việc chấm điểm, PV tiếp theo sẽ được cung cấp. Do đó, các phương pháp này có sự khác biệt về thời gian bình chọn của từng nhóm. Do phương pháp SAMVIQ cho phép người xem kiểm soát việc phát lại của mỗi PV, nhưng trong nghiên cứu này hạn chế thời gian đánh giá tối đa là 4 phút

cho tất cả 8 PV. Mỗi nhóm nhỏ được đo tổng thời gian đánh giá cho tất cả 42 PV bằng từng phương pháp, kết quả được tóm tắt trong Bảng 6. Độ dài tổng thời gian đánh giá cho từng phương pháp từ ngắn nhất đến dài nhất là; ACR5, ACR11, DSIS, SAMVIQ, và DSCQS. Đối với ACR11, phương pháp trình chiếu video giống như ACR5, nhưng tổng thời gian đánh giá lại dài hơn ACR5. Người tham gia cho rằng ACR11 tốn thời gian nhiều hơn cho việc đánh giá vì phương pháp này có nhiều cấp độ hơn ACR5. Tổng thời gian đánh giá của SAMVIQ nhiều hơn DSIS, điều này do SAMVIQ cho phép xem và bình chọn lặp lại của một PVS.

B. Đánh giá sự đơn giản và chính xác

Sau khi hoàn thành thí nghiệm, mỗi người tham gia được đưa một bảng câu hỏi khảo sát về mức độ dễ dàng của việc đánh giá cho mỗi phương pháp đánh giá bằng cách sử dụng thang đo năm mức độ (1: khó, 2: hơi khó, 3: trung bình, 4: hơi dễ, 5: dễ). Điểm trung bình về độ dễ dàng của việc đánh giá cho mỗi phương pháp được hiển thị dưới tên của từng phương pháp trên trục ngang. Điểm trung bình về mức độ dễ dàng của đánh giá đều trên 3, nghĩa là "trung bình" cho bất kỳ phương pháp nào. Kết quả cho thấy, độ dễ dàng của việc đánh giá bằng ACR11 thấp hơn rất nhiều so với ACR5. ACR11 có nhiều mức độ chất lượng video hơn ACR5. Ví dụ, trong ACR5, tốt là mức 4, nhưng trong ACR11, mức 6 và 8 được lựa chọn, vậy có thể xác định mức 7, là mức bắt đầu đạt tốt. Do đó, DSCQS và SAMVIQ, có nhiều mức độ chất lượng video, thấp hơn so với ACR5 và DSIS về độ dễ dàng của đánh giá, giống như ACR11.

V. KẾT LUẬN

Đặc điểm đánh giá chất lượng của DSCQS, DSIS, ACR5, ACR5-HR, ACR11, ACR11-HR, SAMVIQ và SAMVIQHR cho các thể loại ảnh video đã được so sánh. Bài báo đã đưa ra được nhận định rằng hệ số tương quan và xếp hạng hệ số tương quan của điểm số giữa các cặp phương pháp trong 8 phương pháp này đều cao, cho thấy rằng việc lựa chọn thang đo đánh giá sẽ không ảnh hưởng tới kết quả đánh giá chất lượng ảnh video của từng phương pháp. Về thống kê về độ tin cậy, DSIS, ACR5 và SAMVIQ có giá trị trung bình chuẩn hóa của khoảng tin cậy 95% của MOS nhỏ hơn các phương pháp khác. Hơn nữa, kết quả khảo sát của người tham gia cho thấy ACR5 là phương pháp phù hợp từ góc độ tổng thời gian đánh giá và sự dễ dàng trong việc đánh giá. Qua đó, ACR5 được coi là phương pháp phù hợp nhất để đánh giá chất lượng các thể loại ảnh video tại thời điểm hiện nay.

TÀI LIỆU THAM KHẢO

1. ITU-R Rec. BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," June 2002.
2. ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications," April 2008.
3. ITU-R Rec. BT.1788, "Methodology for the subjective assessment of video quality in multimedia applications," January 2007.

4. P. Coriveau, C. Gojmerac, B. Hughes, and L. Stelmach, "All subjective scales are not created equal: The effects of context on different scales," Signal Processing, Vol. 77, August 1999
5. ITU-R Document 6Q/131-E, "Technical report: Comparison of DSCQS and ACR," October 2005.
6. ITU-R Document 6Q/208-E, "Report on experiment of new subjective video quality metrics SAMVIQ for mobile video," April 2007.
7. M. Brotherton, Q. Huynh-Thu, D. Hands, and K. Brunnstrom, "Subjective multimedia quality assessment," IEICE Trans. on Fundamentals, No. 11, Vol. E89-A, November 2006.
8. S. Pechard, R. P'epion, and P. Le Callet, "Suitable methodology in subjective video quality assessment: A resolution dependent paradigm," IMQA 2008, September 2008.
9. Kaggle Dataset (available online), <https://www.kaggle.com/datasets>.
10. M. Řeřábek and T. Ebrahimi, "New Light Field Image Dataset," 8th International Workshop on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal, 2016.
11. VISlab dataset, (available online), <https://ivi.fnwi.uva.nl/vislab/>
12. Z. K. Kalazhokov, Y.T. Makoveichuk, "Seminar on Computer Science and Information Technologies", 2023 - ieeexplore.ieee.org
13. E. Serrano, M. Molina, D Manrique, et al, "Intelligent Environments", 2017 - ebooks.iospress.nl

SUBJECTIVE QUALITY ASSESSMENT METHODS: A COMPREHENSIVE PERFORMANCE STUDY

Abstract: For the current research area on video image quality, there are various subjective evaluation methods depending on the purpose. Subjective evaluation methods are recommended by ITU-T and ITU-R. However, surveys indicate a lack of comprehensive studies on selecting the most appropriate and accurate subjective evaluation method. Therefore, this paper provides a comprehensive study based on a comparison of the most common current subjective evaluation methods. In this paper, eight subjective evaluation methods are examined, including the Double-Stimulus Continuous Quality-Scale (DSCQS), the Double-Stimulus Impairment Scale (DSIS), Absolute Category Rating with a 5-point scale (ACR5), Absolute Category Rating with a 5-point scale with hidden reference (ACR5-HR), ACR with an 11-level scale (ACR11), ACR11 with hidden reference (ACR11-HR), Subjective Assessment of Multimedia Video Quality (SAMVIQ), and SAMVIQ with hidden reference (SAMVIQ-HR). These methods are compared and tested in various scenarios to determine each method's applicability for image and video evaluation. Assessment aspects include the total time and difficulty/ease of evaluation for participants, as well as the characteristics of different scales and the statistical reliability of each method. The results indicate the correlation coefficient and distribution of correlation in mean opinion scores (MOS) are high among the methods. Furthermore, the results clearly demonstrate that the statistical reliability parameters of evaluation methods like DSIS, ACR5, and SAMVIQ outperform others.

Additionally, participant feedback shows that ACR5 is the most suitable method based on criteria of total evaluation time and ease of assessment.

Keywords: Subjective evaluation method, MOS, image and video quality, user experience quality (QoE).



Phí Công Huy tốt nghiệp đại học Oxford Brookes, UK; lấy bằng thạc sĩ tại Monash, Úc, lấy bằng tiến sĩ tại đại học công nghệ Sydney UTS. Hiện đang công tác tại Khoa Đa phương tiện, Học viện Công nghệ Bưu chính Viễn thông, nghiên cứu chuyên sâu các lĩnh vực liên quan tới xử lý ảnh, video, kiểm soát chất lượng dữ liệu ảnh Light

Field.