ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

PHÙNG THÁI THIÊN TRANG

TĂNG CƯỜNG KHẢ NĂNG HIỂU ẢNH CỦA HỆ THỐNG TRUY VẤN ẢNH

Ngành: Khoa học Máy tính
Mã số ngành: 62480101

TÓM TẮT LUẬN ÁN TIẾN SĨ
KHOA HỌC MÁY TÍNH

Tp. Hồ Chí Minh năm 2024

Công trình được hoàn thành tại:
Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM

Người hướng dẫn khoa học:
1. HDC: PGS.TS. Lý Quốc Ngọc
2. HDP: PGS.TS. Fukuzawa Masayuki

Phản biện 1: PGS.TS. Huỳnh Trung Hiếu

Phản biện 2: PGS.TS. Lê Đình Duy

Phản biện 3: TS. Hà Việt Uyên Sinh

Phản biện độc lập 1: PGS.TS. Huỳnh Trung Hiếu

Phản biện độc lập 2: TS. Ngô Quốc Việt

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án cấp Cơ sở đào tạo họp tại
Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM,
vào hồi … giờ ……, ngày …….. tháng …….. năm …….

Có thể tìm hiểu luận án tại thư viện:
1. Thư viện Tổng hợp Quốc gia Tp.HCM
2. Thư viện trường Đại học Khoa học Tự nhiên, ĐHQG-HCM
3. Thư viện trung tâm ĐHQG HCM

MỤC LỤC

Chương 1. MỞ ĐẦU

1.1 Bối cảnh chung

Hiện nay, Trí tuệ nhân tạo đóng vai trò quan trọng trong xây dựng các ứng dụng thông minh, mang lại hiệu quả ngày càng cao cho đời sống con người như trong học tập, làm việc, chăm sóc sức khoẻ .v.v. Trong bối cảnh đó, nhu cầu xây dựng hệ thống thị giác thông minh ngày càng cấp thiết.

Hiện nay, truy vấn ảnh có hai xu hướng trong các ứng dụng thực tế: (1) Truy vấn ảnh tổng quát, thường thấy trong công cụ tìm kiếm như Google, Bing, v.v, và (2) Truy vấn ảnh đặc thù, như trong truy vấn xe, biển số xe trong giao thông minh, truy vấn người, mặt người trong giám sát an ninh, truy vấn trang phục trong thương mại điện tử.

Để tăng cường tính *thông minh* cho hệ thống truy vấn ảnh, luận án tập trung vào việc tăng cường khả năng hiểu ảnh của hệ thống truy vấn ảnh. Việc này được thực hiện dựa trên việc khai thác *biểu diễn tri thức, các mô hình học sâu, các mô hình học sâu đa phương thức* nhằm: (1) tăng cường khả năng hiểu ảnh trong giai đoạn tổ chức dữ liệu bằng cách ***rút trích đặc trưng mang tính ngữ nghĩa cao*** nhờ vào đặc trưng học sâu và học thuộc tính, (2) tăng cường khả năng hiểu ảnh trong giai đoạn truy vấn bằng cách đưa vào cách ***biểu diễn câu truy vấn linh hoạt*** nhờ vào văn bản và hình ảnh, đồng thời ***điều hướng truy vấn*** dựa vào phả hệ tri thức.

Tóm lại: mục đích chính của luận án "Tăng cường khả năng hiểu ảnh của hệ thống truy vấn ảnh" là xây dựng Hệ thống truy vấn ảnh thông minh hỗ trợ truy vấn tập dữ liệu tổng quát, tập dữ liệu đặc thù với công cụ cốt lõi là trí tuệ nhân tạo, dựa trên phả hệ tri thức đối tượng, mô hình học thuộc tính, mô hình đa phương thức, ở mức toàn cục và mức chi tiết.

1.2 Các bài toán con

Luận án chia thành các bài toán con sau:

- Bài toán con 1: truy vấn ảnh dựa vào đặc trưng học sâu
- Bài toán con 2: truy vấn ảnh dựa vào phả hệ tri thức và thuộc tính đối tượng

- Bài toán con 3: truy vấn ảnh dựa vào đa phương thức kết hợp văn bản và hình ảnh

1.3 Phạm vi của bài toán

Luận án nhắm đến các đối tượng và thuộc tính đối tượng cần khảo sát, thực nghiệm là các đối tượng như *xe*, *người đi đường, mặt người và trang phục*.

1.4 Các đóng góp của luận án

- Xây dựng một phả hệ tri thức thuộc tính đối tượng hỗ trợ mô đun học thuộc tính và truy vấn. phả hệ tri thức đảm nhận vai trò lưu trữ tri thức có sẵn, phân cấp dữ liệu nhằm tái sử dụng tri thức và điều hướng học thuộc tính và truy vấn.
- Xây dựng một mô hình học thuộc tính đối tượng giúp hệ thống hiểu ảnh mở mức tinh, chi tiết và vì thế giúp tăng độ chính xác và thông minh hơn vì phát hiện ra thuộc tính đối tượng sau đó dễ dàng tìm thấy ảnh theo thuộc tính đối tượng.
- Xây dựng hệ thống truy vấn ảnh thông minh: (1) dựa vào đặc trưng học sâu, (2) dựa vào học thuộc tính và phả hệ tri thức, (3) kết hợp văn bản và hình ảnh trong truy vấn ảnh.

Từ kết quả trên, luận án đạt được một số đóng góp sau đây:

1.4.1 Về mặt lý thuyết

(1) Đóng góp thứ nhất: Tăng cường khả năng hiểu ảnh của hệ thống truy vấn ảnh dựa vào đặc trưng học sâu

(2) Đóng góp thứ hai: Tăng cường khả năng hiểu ảnh của hệ thống truy vấn ảnh nhờ vào mô hình nhận dạng đối tượng ở mức thuộc tính.

(3) Đóng góp thứ ba: Tăng cường khả năng hiểu ảnh của hệ thống truy vấn ảnh nhờ vào mô hình đa phương thức kết hợp ảnh và câu văn bản mô tả.

1.4.2 Về mặt ứng dụng

(4) Cải thiện độ chính xác, tính tiện dụng và thông minh

(5) Linh hoạt đầu vào truy vấn, giảm chi phí, tăng hiệu suất

(6) Hỗ trợ truy vấn y phục, hỗ trợ truy vấn người dựa vào thuộc tính mặt người, y phục

Tóm lại: Luận án đã xây dựng một hệ thống truy vấn ảnh mang lại các đóng góp chính như cải thiện *độ chính xác, tiện dụng, thông minh, tái sử dụng tri thức có sẵn.*

1.5 Tổ chức luận án

Luận án được tổ chức thành 5 chương như sau:

- Chương 1. Giới thiệu
- Chương 2. Cơ sở lý thuyết và các nghiên cứu liên quan
- Chương 3. Tăng cường khả năng hiểu ảnh của hệ thống truy vấn ảnh
- Chương 4. Thực nghiệm và đánh giá
- Chương 5. Kết luận và hướng phát triển
- Phần phụ lục

Chương 2. CÁC CÔNG TRÌNH NGHIÊN CỨU LIÊN QUAN

2.1 Truy vấn ảnh dựa vào thuộc tính

❖ *Mô hình học thuộc tính*

Để hiểu rõ ảnh ở mức mịn hơn, cụ thể là hiểu đối tượng ở mức thuộc tính, các hệ truy vấn sử dụng kết quả học thuộc tính đối tượng để hỗ trợ truy vấn ở mức mịn hơn. Điều này cũng thể hiện hệ truy vấn thông minh hơn vì có thể truy vấn ảnh qua mô tả thuộc tính đối tượng trên ảnh. Một số công trình điển hình về học thuộc tính như: nhóm tác giả Zisserman và Vittorio Ferrari [41], nhóm tác giả Adriana Kovashka và Kristen Grauman trong [44][45]. Các nghiên cứu khác cũng sử dụng thuộc tính đối tượng hiệu quả trong truy vấn ảnh như [24] [37] với dữ liệu lớn đa dạng [38]. Các công trình điển hình về sử dụng mô hình học sâu để học thuộc tính gồm PANDA [46], Sun Attribute Database for scene [47], DEEP-CARVING [48].

❖ *Truy vấn ảnh người đi dường dựa vào thuộc tính*

Những năm đây, các công trình nghiên cứu về nhận dạng ảnh người đi đường (pedestrian) và tái định danh người phát triển mạnh. Các công trình điển hình như [66] [67][68][69].

2.2 Truy vấn ảnh dựa vào phả hệ tri thức

Chức năng của Phả hệ tri thức là dùng để mô tả, lưu trữ và suy diễn thông tin thuộc một phạm vi nhất định, ngoài ra có thể mở rộng và tái sử dụng phả hệ tri thức dễ dàng. Một số công trình điển hình sử dụng đề xuất phả hệ tri

thức và sử dụng hiệu quả như Phả hệ tri thức đối tượng [34], Phả hệ tri thức khái niệm thị giác [77], Phả hệ tri thức đối tượng mặt người FAO [78], đối tượng thời trang [72], người đi đường [71]. Luận án kế thừa các phả hệ tri thức của công trình [78] [71] [72] để xây dựng một phả hệ tri thức mới lưu trữ tri thức thuộc tính đối tượng nhằm hỗ trợ hệ truy vấn ảnh về mặt ngữ nghĩa. Nội dung cụ thể sẽ được trình bày chi tiết trong Chương 3.

2.3 Truy vấn ảnh dựa vào văn bản và hình ảnh

Sử dụng học sâu để xây dựng hệ thống truy vấn ảnh dựa vào văn bản là cách làm mới, đặc biệt sử dụng mạng transfomer và các biến thể của nó, một số phương pháp điển hình như CLIP [87] [88], Truy vấn ảnh đa phương thức tích hợp văn bản và hình ảnh của nhóm tác giả Ding Jiang [88].

Chương 3. TĂNG CƯỜNG KHẢ NĂNG HIỂU ẢNH CỦA HỆ THỐNG TRUY VẤN ẢNH

3.1 Truy vấn ảnh dựa vào đặc trưng học sâu

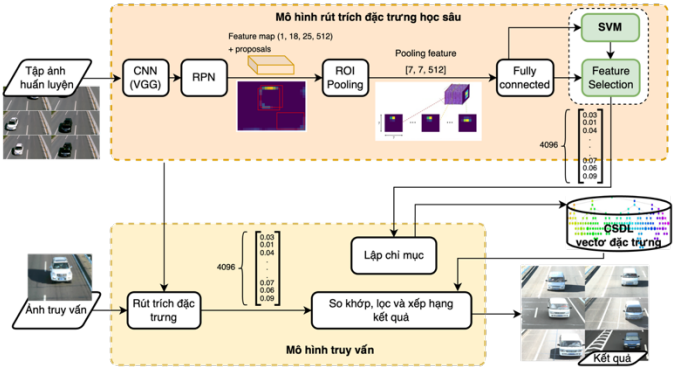
Trong phần này, luận án trình bày đề xuất một hệ thống truy vấn ảnh dựa vào đặc trưng học sâu. Cụ thể, luận án sử dụng mạng học sâu Faster R-CNN [89] kết hợp với SVM để rút trích đặc trưng ảnh. Sau đó sử dụng phương pháp ANN (Approximate Nearest Neighbors), cụ thể là công cụ Annoy [28] để lập chỉ mục các vectơ đặc trưng của tập dữ liệu và hỗ trợ tìm kiếm nhanh trong bước truy vấn ảnh. Để minh chứng cho sự hiệu quả của hệ thống, luận án thực nghiệm với ảnh phương tiện giao thông. Sau đây, luận án trình bày chi tiết hệ thống truy vấn ảnh gồm hai mô hình chính là: mô hình rút trích đặc trưng và mô hình truy vấn ảnh.

3.1.1 Mô hình rút trích đặc trưng ảnh dựa vào mạng học sâu

Trong mô hình này, luận án đã sử dụng xây dựng mô hình rút trích đặc trưng dựa vào mạng Faster R-CNN [89] kết hợp với SVM. Đầu tiên dùng Faster R-CNN để rút trích ra véctơ đặc trưng, sau đó, sử dụng SVM để phân loại và lựa chọn đặc trưng kết quả.

Mô đun rút trích đặc trưng ảnh sẽ trả về hai tham số: {lớp, vectơ đặc trưng}, một là vectơ đặc trưng ảnh có độ dài 4096, hai là ảnh thuộc lớp nào. Hai tham số này sẽ làm đầu vào cho giai đoạn truy vấn.

3.1.2 Mô hình truy vấn ảnh dựa vào đặc trưng học sâu



Hình 3.5: Mô hình truy vấn ảnh dựa vào mạng học sâu

3.2 Truy vấn ảnh dựa vào đặc trưng học sâu và học thuộc tính

3.2.1 Xây dựng mô hình học thuộc tính

Trong phần này, luận án trình bày đề xuất mô hình học thuộc tính đối tượng dựa vào đặc trưng học sâu. Đây là một mô đun quan trọng trong hệ thống, giúp hệ thống truy vấn ảnh có khả năng hiểu ảnh ở mức thuộc tính theo hướng tiếp cận từ thô đến tinh, hệ thống cho phép tìm ảnh theo thuộc tính để thể hiện khả năng hiểu ảnh tốt hơn vì có thể hiểu ảnh ở mức chi tiết.

Các bước thực hiện:

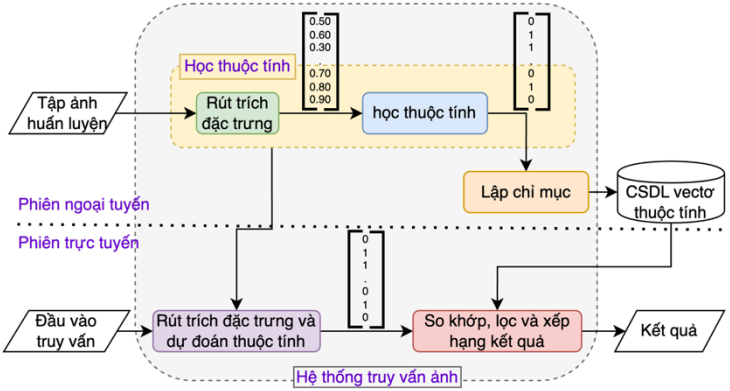
- Bước 1: Tiền xử lý
- Bước 2: Rút trích đặc trưng
- Bước 3: Huấn luyện thuộc tính
- Bước 4: trả về mô hình học thuộc tính

❖ *Rút trích đặc trưng ảnh:* Sử dụng mô hình mạng học sâu EfficientNet của nhóm tác giả Mingxing Tan và Lê Viết Quốc [91] để rút trích đặc trưng ảnh với độ dài vectơ đặc trưng là 2048.

❖ *Giai đoạn huấn luyện*: Luận án sử dụng bộ phân lớp tuyến tính Linear

Kết quả phần thực nghiệm về học thuộc tính được trình bày trong Chương 4, mục thực nghiệm 2

3.2.2 Truy vấn ảnh dựa vào thuộc tính



Hình 3.9 Mô hình truy vấn ảnh dựa vào thuộc tính

3.2.3 Tiểu kết 1

Trong phần trên này, luận án đã đưa ra hai phương án xây dựng hệ thống truy vấn ảnh là truy vấn ảnh dựa vào đặc trưng học sâu và truy vấn ảnh dựa vào đặc trưng toàn cục và học thuộc tính. Hai phương án này đáp ứng một lớp rộng phạm vi dữ liệu ảnh trong thực tế từ dữ liệu tổng quát đến dữ liệu đặc thù.

3.3 Xây dựng phả hệ tri thức thuộc tính đối tượng

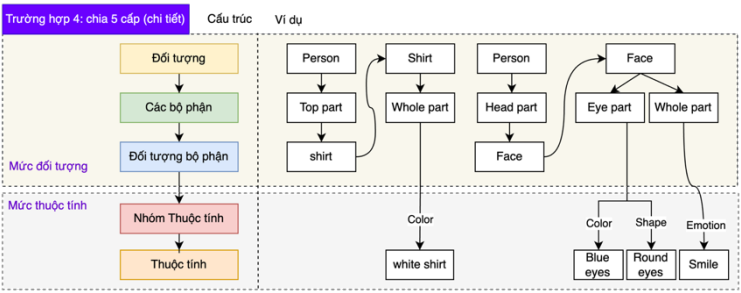
3.3.1 Phả hệ tri thức thuộc tính đối tượng

Định nghĩa: Phả hệ tri thức thuộc tính đối tượng (Object Attribute Ontology-OAO) bao gồm năm thành phần, ký hiệu **OAO**(C, P, Re, Ru, I) như sau với:

- Các khái niệm (C-concepts) trong Phả hệ tri thức thuộc tính đối tượng này sẽ bao gồm (1) khái niệm đối tượng, (2) khái niệm bộ phận của đối

tượng, (3) khái niệm thuộc tính đối tượng, (4) khái niệm nhóm thuộc tính, (5) Khái niệm thị giác (visual concepts).

- Phân cấp khái niệm Tùy theo mỗi đối tượng mà có thể tổ chức phân cấp các khái niệm cho phù hợp với đối tượng cụ thể. Luận án chia thành bốn (04) trường hợp phân cấp khái niệm cho các đối tượng như sau
  - Trường hợp 1: chia 2 cấp, trường hợp này phù hợp cho các đối tượng đơn giản chỉ có một bộ phận như quả bóng, quả táo (trái cây).
  - Trường hợp 2: chia 3 cấp, trường hợp này phù hợp với các đối tượng chia thành nhiều nhóm thuộc tính như nhóm khái niệm thị giác màu, vân, dáng, kích thước theo phả hệ tri thức thị giác.
  - Trường hợp 3: chia 4 cấp, trường hợp này phù hợp với các đối tượng chia nhiều bộ phận như mặt người hoặc áo (có một bộ phận wholepart).
  - Trường hợp 4: chia 5 cấp, trường hợp này phù hợp với các đối tượng phức tạp, chia nhiều bộ phận và trong mỗi bộ phận lại có chứa đối tượng khác như con người chia 4 bộ phận gồm phần đầu (head) có chứa đối tượng mặt người, phần thân trên (top) có chứa đối tượng áo, …



Hình 3.16. Tổ chức 5 cấp khái niệm (chi tiết)

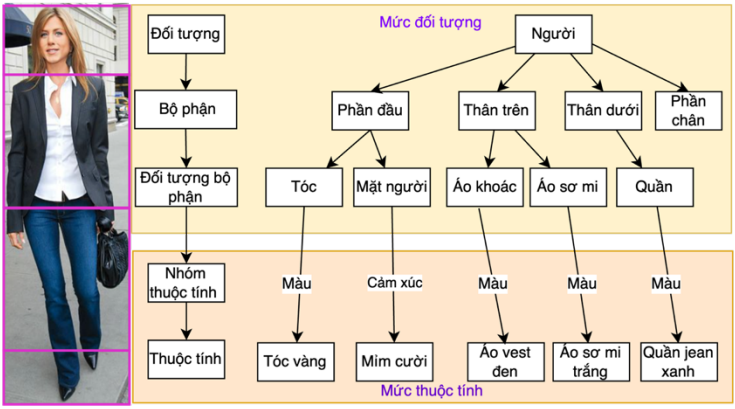
- Mối quan hệ: Các khái niệm trong phả hệ tri thức luôn có mối quan hệ lẫn nhau như quan hệ giữa các đối tượng hoặc thuộc tính: "là con", "có chứa", "là bộ phận", ….

- Tập luật: Tập luật dùng để suy diễn tri thức dựa trên luật. Các tri thức có sẵn và suy diễn được sẽ biểu diễn thành luật. Nhờ tập luật, các tri thức suy diễn được sẽ được phát hiện mà không cần học. Từ đó giúp tiết kiệm chi phí huấn luyện. Ví dụ về luật :Mặt người và "có râu"→ đàn ông,

Mỗi luật bao gồm (1) đầu vào, (2) điều kiện, (3) đầu ra.

$$\text{đầu vào} \xrightarrow{\text{điều kiện}} \text{đầu ra}$$

3.3.2 Xây dựng phả hệ tri thức thuộc tính đối tượng người



Hình 3.21. Ví dụ trích cây phân cấp thuộc tính đối tượng người

Trong phần này, luận án trình bày đề xuất một phả hệ tri thức thuộc tính đối tượng người (HAO-Human Attribute Ontology). Dựa vào phả hệ tri thức thuộc tính đối tượng **OAO** để xây dựng phả hệ tri thức thuộc tính đối tượng người **HAO**. Luận án tham khảo từ các phả hệ tri thức PAO [71], FashionOA [72], FAO [78] để xây dựng nên HAO.

Các khái niệm trong HAO sẽ bao gồm, người, các bộ phận (đầu, thân trên, thân dưới, chân, v.v), các đối tượng bộ phận như áo, quần, váy, đầm, nón mũ, các thuộc tính như áo đỏ, quần xanh, v.v.

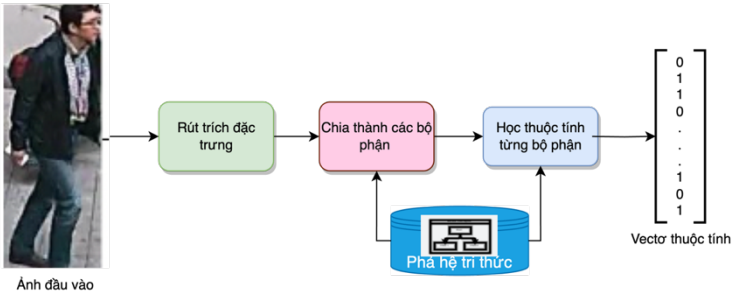
Luận án thực hiện khảo sát các tập dữ liệu về người sẽ dùng cho thực nghiệm bao gồm: PETA 1: 35 thuộc tính và 105 thuộc tính, PA100k: 26 thuộc tính,

RAP: 92 thuộc tính, RAPv2:152 thuộc tính, chi tiết về các tập dữ liệu được trình bày trong phần phụ lục A

Cây phân cấp khái niệm chia thành 5 cấp được biểu diễn theo hình dưới đây:

3.4 Xây dựng mô hình học thuộc tính dựa vào phả hệ tri thức

3.4.1 Mô hình Học thuộc tính



Hình 3.29 Mô hình học thuộc tính đa nhiệm dựa vào đặc trưng cục bộ

Các bước thực hiện:

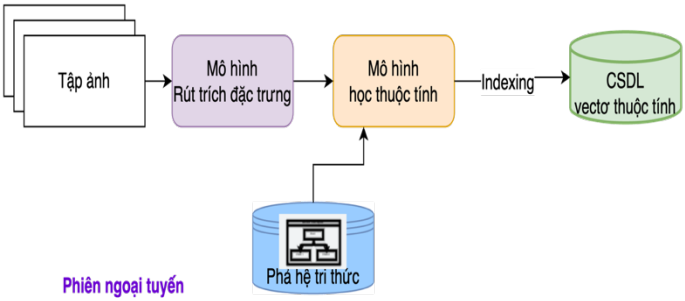
- Bước 1: Tiền xử lý
- Bước 2: rút trích đặc trưng
- Bước 3: chia theo các bộ phận
- Bước 4: huấn luyện thuộc tính trên mỗi bộ phận
- Bước 5: xử lý và trả về vectơ thuộc tính

3.5 Xây dựng hệ thống truy vấn ảnh dựa vào học thuộc tính và phả hệ tri thức

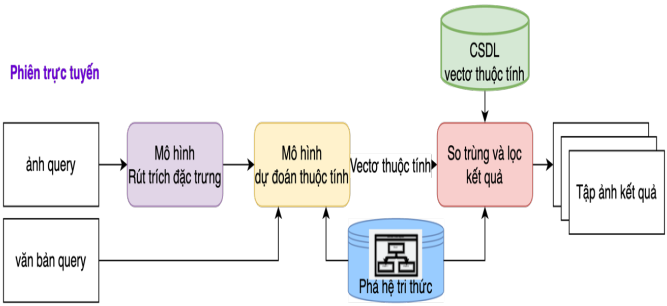
3.5.1 Mô tả hệ thống truy vấn ảnh

Hệ thống gồm có hai phiên: phiên ngoại tuyến và phiên trực tuyến. Mỗi phiên đảm nhận một số nhiệm vụ quan trọng. Phiên ngoại tuyến bao gồm các nhiệm vụ như (1) xây dựng phả hệ tri thức thuộc tính đối tượng, (2) Xây dựng dựng mô hình học thuộc tính, (3) Lập chỉ mục và lưu vào cơ sở dữ liệu chuẩn bị phục vụ bước truy vấn và (4) chức năng dự đoán thuộc tính chuẩn bị cho bước trực tuyến. Phiên trực tuyến gồm (1) xử lý dữ liệu đầu vào truy vấn

(query), (2) so khớp và lọc ra các ảnh tương đồng với query và (3) cuối cùng là sắp xếp kết quả.



Hình 3.34: Phiên ngoại tuyến

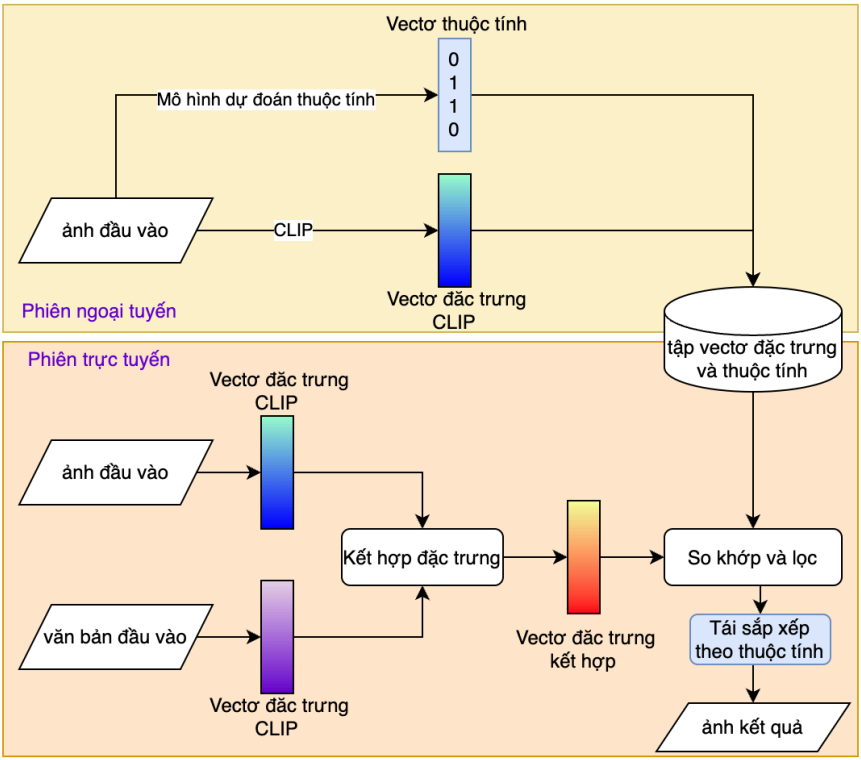


Hình 3.35: Phiên trực tuyến

3.6 Tích hợp văn bản và hình ảnh trong truy vấn ảnh

Luận án sử dụng mô hình CLIP [87] và mô hình học thuộc tính để xây dựng mô hình truy vấn ảnh như hình 3.42

Thực hiện kết hợp hai vectơ đặc trưng theo phương pháp tính tích 2 vectơ: text_feature ⊗ image_feature để hình thành 1 vectơ đại diện cho query



Hình 3.42. Truy vấn ảnh dựa vào thuộc tính và tích hợp văn bản và hình ảnh

3.7 Kết chương

Hệ thống truy vấn ảnh của luận án hoàn thành với nhiệm vụ tăng cường khả năng hiểu ảnh của hệ thống truy vấn ảnh được thể hiện qua kết quả của (1) hệ thống truy vấn ảnh dựa vào đặc trưng học sâu, (2) hệ thống truy vấn ảnh dựa vào học thuộc tính và phả hệ tri thức thuộc tính đối tượng, và (3) tích hợp văn bản và hình ảnh trong truy vấn ảnh. Từ kết quả của các đề xuất trên, hệ thống truy vấn ảnh thể hiện sự tăng cường khả năng hiểu ảnh và tính thông minh, linh hoạt.

Chương 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

4.1 Thực nghiệm 1: Hệ thống truy vấn ảnh dựa vào đặc

trưng học sâu

Thực nghiệm hệ thống truy vấn ảnh phương tiện giao thông

4.1.1 Mô tả thực nghiệm

Trong phần này, luận án đánh giá phương pháp đề xuất bằng cách thực nghiệm trên tập dữ liệu BIT-Vehicle [95]. Tập dữ liệu chứa 9652 hình ảnh phương tiện giao thông với kích thước 1600x1200 và 1920x1080, gồm sáu lớp (classes) với xe buýt (558), microbus (883), xe tải nhỏ (476), sedan (5922), SUV (1392) và xe tải (822). Hình ảnh mẫu và số lượng ảnh trong mỗi nhóm lớp được trình sau đây:

4.1.2 Kết quả phân loại phương tiện giao thông

(ký hiệu P: Precision, R: Recall)

BẢNG 5.2. KẾT QUẢ PHÂN LOẠI ẢNH TRÊN TẬP BIT-VEHICLE (CHI TIẾT)

Các lớp xe	AlexNet		Faster-RCNN		Luận án :Faster-R-CNN +SVM	
	P	***R***	***P***	***R***	***P***	***R***
Bus	99.04	98.10	100.00	97.14	100	100
Microbus	86.18	81.37	96.60	88.20	95.71	96.89
Minivan	86.11	88.57	90.99	96.19	97.14	97.14
SUV	83.00	83.33	84.62	98.37	91.51	96.34
Sedan	96.95	97.37	99.82	95.88	99.38	98.07
Truck	94.80	95.35	95.45	97.67	98.26	98.26
Average	91.01	90.68	94.58	95.58	97.00	97.78

BẢNG 5.3. KẾT QUẢ PHÂN LOẠI ẢNH TRÊN TẬP BIT-VEHICLE (TRUNG BÌNH)

Phương pháp	Precision Avg	Recall Avg	Accuracy Avg
AlexNet	91.01	90.68	93.66
Faster-RCNN	94.58	95.59	95.86
Luận án: Faster-R-CNN +SVM	97.00	97.78	97.82
Tiny-YOLO [96]	**97.90**	**99.60**	-

4.1.3 Kết quả truy vấn ảnh

Kết quả truy vấn được trình bày trong bảng dưới đây. Với cùng một hệ thống cài đặt và dữ liệu dành cho huấn luyện và kiểm tra, luận án đã đạt được mAP

tốt nhất trong truy vấn ảnh so với phương pháp Alexnet và Faster R-CNN. (Ký hiệu P: precision, R: Recall)

BẢNG 5.4. KẾT QUẢ TRUY VẤN ẢNH TRÊN TẬP BIT-VEHICLE (CHI TIẾT)

Các lớp xe	AlexNet		Faster-RCNN		Luận án:Faster-R-CNN +SVM+ANN	
	P	***R***	***P***	***R***	***P***	***R***
Bus	6.76	75.78	99.45	77.08	**99.95**	82.49
Microbus	8.64	63.79	88.76	54.59	82.34	61.22
Minivan	4.53	63.40	87.76	79.09	89.87	78.11
SUV	14.51	65.37	83.43	76.50	81.68	75.14
Sedan	58.07	64.73	99.39	87.49	99.25	**88.88**
Truck	9.56	74.00	96.46	81.74	96.52	82.75
Average	17.01	67.85	92.54	76.08	91.60	78.10

BẢNG 5.5. KẾT QUẢ TRUY VẤN ẢNH TRÊN TẬP BIT-VEHICLE (TRUNG BÌNH)

Phương pháp	AlexNet	Faster-RCNN	Luận án :Faster-RCNN+SVM+ANN
MAP	46%	94%	**95.23%**

4.2 Thực nghiệm 2: Hệ thống truy vấn ảnh dựa vào học thuộc tính và đặc trưng học sâu toàn cục

4.2.1 Tập dữ liệu

Luận án sử dụng các tập dữ liệu theo về người đi đường như PETA [97], PA100k [98], RAP [99], and RAP2 [100]. Chi tiết về các tập dữ liệu được trình bày trong phụ lục A). Thông tin số lượng ảnh và thuộc tính được trình bày trong bản dưới đây:

Phân bổ số lượng ảnh cho tập huấn luyện và kiểm thử (train, val, test)

BẢNG 4.7. CÁC TẬP TRAIN, VAL, TEST

Tập dữ liệu	Số ảnh	Số thuộc tính	Tập train	Tập val	Tập test
Pa100k	100,000	26	80,000	10,000	10,000
PETA	19,000	35	9,500	1,900	7,600
RAP	41,585	92	33,268		8,317
RAP2	84,928	152	50,957	16,986	16,985

Phương pháp đánh giá: Luận án sử dụng các phương pháp đánh giá phổ biến trình bày trong phụ lục B.

4.2.2 Thực nghiệm mô hình học thuộc tính dựa vào đặc trưng học sâu

toàn cục

Luận án sử dụng EfficientNetB5 [91] để rút trích đặc trưng. Các ảnh được chỉnh lại kích thước 224x224 trước khi gởi đến EfficientNet. Đầu ra của bước rút trích đặc trưng sẽ trả về vectơ có độ dài là 2048 sẽ làm đầu vào cho giai đoạn phân loại theo thuộc tính. Luận án sử dụng bộ phân lớp tuyến tính Linear của thư viện Torch cho bước phân loại thuộc tính.

- Kết quả thực nghiệm mô hình học thuộc tính dựa vào đặc trưng học sâu toàn cục thể hiện trong các bảng 4.10, 4.11, 4.12, 4.13 ở dòng dữ liệu không sử dụng phả hệ tri thức.

4.3 Thực nghiệm 3: Xây dựng hệ thống truy vấn ảnh dựa vào học thuộc tính và phả hệ tri thức

4.3.1 Thực nghiệm xây dựng phả hệ tri thức thuộc tính đối tượng

Luận án thực hiện tổ chức tri thức cho các phả hệ tri thức về người, mặt người và thời trang. Mỗi đối tượng sẽ có một hoặc một số bộ phận, một số nhóm thuộc tính và các thuộc tính của đối tượng. Số khái niệm về đối tượng và đặc biệt và thuộc tính về người, mặt người và thời trang là rất nhiều. Tuy nhiên, do thời gian có hạn nên luận án chỉ thu thập được một số khái niệm cơ bản như trong thực nghiệm với HAO gồm 235 thuộc tính từ các tập dữ liệu PETA, Pa100k, RAP, RAPv2.

4.3.2 Thực nghiệm mô hình học thuộc tính dựa vào phả hệ tri thức, học đa nhiệm và đặc trưng cục bộ

Kết quả có sử dụng phả hệ tri thức và không sử dụng phả hệ tri thức

Kết quả học thuộc tính trên tập PETA:

BẢNG 4.10. KẾT QUẢ HỌC THUỘC TÍNH TRÊN TẬP PETA

Công trình	mA	Accuracy	Precision	Recall	F1-score
Zichang Tan [101]	84.88	79.46	87.42	86.33	86.87
HydraPlus-Net [98]	81.77	76.13	84.92	83.24	84.07
ALM [102]	86.3	79.52	85.65	88.09	86.85
Strong Baseline [92]	85.11	79.14	86.99	86.33	86.09
Jian Jia 2021 [103]	86.52	78.95	86.02	87.12	86.99
feature pyramid [104]	87.69	**81.2**	87.59	**89.2**	88.32

Công trình	mA	Accuracy	Precision	Recall	F1-score
DAFL[105]	87.07				86.4
JLAC [106]	86.96	80.38	87.81	87.09	87.45
Rein-PAR[107]	85.51	78.45	84.08	88.77	85.91
Dai[108]	88.24	79.14	**88.79**	84.7	86.7
UPAR [109]	**88.40**				**89.9**
Ours:AL w/0 HAO	85.50	79.50	87.32	86.45	86.60
Ours:AL + HAO	85.41	80.17	88.09	86.59	87.07

Kết quả học thuộc tính trên tập Pa100k:

BẢNG 4.11. KẾT QUẢ HỌC THUỘC TÍNH TRÊN TẬP PA100K

Công trình	mA	Accuracy	Precision	Recall	F1-score
Zichang Tan [101]	81.61	78.89	86.83	87.73	87.27
HydraPlus-Net [98]	74.21	72.19	82.97	82.09	82.53
ALM [102]	80.68	77.08	84.21	88.84	86.46
Strong Baseline [92]	79.38	78.56	89.41	84.78	86.25
Jian Jia 2021 [103]	81.87	78.86	85.98	89.1	86.87
feature pyramid [104]	81.45	79.13	86.24	**89.46**	87.94
DAFL[105]	83.54				88.09
JLAC [106]	82.31	79.47	87.45	87.77	87.61
Rein-PAR[107]	80.55	77.2	84.76	87.67	85.7
Dai[108]	77.89	79.71	**90.26**	85.37	87.75
UPAR [109]	**84.80**				**90.2**
Ours:AL w/0 HAO	82.49	**79.89**	87.39	88.32	87.49
Ours:AL + HAO	75.46	75.34	88.86	81.34	84.42

Kết quả học thuộc tính trên tập RAP:

BẢNG 4.12. KẾT QUẢ HỌC THUỘC TÍNH TRÊN TẬP RAP

Công trình	mA	Accuracy	Precision	Recall	F1-score
Zichang Tan [101]	81.25	67.91	78.56	81.45	79.98
HydraPlus-Net [98]	76.12	65.39	77.33	78.79	78.05
ALM [102]	81.87	68.17	74.71	86.48	80.16
Strong Baseline [92]	78.48	67.17	82.84	76.25	78.94
Jian Jia 2021 [103]	82.77	68.37	75.05	**87.49**	80.43
feature pyramid [104]	82.37	**69.93**	80.46	87.23	**82.33**
DAFL[105]	**83.72**				80.29
JLAC [106]	83.69	69.15	79.31	82.40	80.82
Rein-PAR[107]	81.67	66.24	73.24	85.80	78.68
Dai[108]	75.09	66.9	**84.27**	79.16	76.46
Ours:AL w/0 HAO	80.98	68.16	79.61	80.53	79.72
Ours:AL with HAO	66.31	61.85	81.82	70.24	75.00

Kết quả học thuộc tính trên tập RAPv2:

BẢNG 4.13. KẾT QUẢ HỌC THUỘC TÍNH TRÊN TẬP RAPV2

Công trình	mA	Accuracy	Precision	Recall	F1-score
UPAR [109]	79.90				**81.0**
Ours:AL w/o HAO	79.19	68.05	79.18	80.88	79.67
Ours:AL with HAO	**80.43**	67.69	78.18	**81.45**	79.44

4.3.3 Các thuộc tính có độ chính xác cao nhất

Ngoài ra, luận án rút trích ra sáu (06) thuộc tính có kết quả cao nhất trong số 35 thuộc tính về độ đo Accuracy trên tập PETA như sau:

BẢNG 4.14. KẾT QUẢ 6 THUỘC TÍNH NỔI TRỘI

Thuộc tính	Accuracy	Precision	Recall	F1-score
lowerBodyCasual	**94.22**	96.13	**97.94**	**97.03**
upperBodyCasual	93.72	95.87	97.66	96.76
personalLarger60	90.19	**98.14**	91.76	94.84
accessoryMuffler	89.97	97.33	92.25	94.72
personalMale	89.16	93.73	94.82	94.27
accessoryNothing	88.80	93.41	94.74	94.07
Average	91.01	95.77	94.86	95.28

So sánh kết quả có **6** thuộc tính đạt kết quả nổi trội về độ đo accuracy so với công trình điển hình DeepMAR [110] như sau:

BẢNG 4.15. SO SÁNH KẾT QUẢ 6 THUỘC TÍNH NỔI TRỘI

Thuộc tính	DeepSAR [110]	DeepMAR [110]	Ours	+/-
lowerBodyCasual	81.60	84.90	**94.22**	+
upperBodyCasual	81.10	84.40	**93.72**	+
personalLarger60	92.00	**94.80**	90.19	-
accessoryMuffler	94.40	**96.10**	89.97	-
personalMale	85.10	**89.90**	89.16	-
accessoryNothing	81.50	85.80	**88.80**	+
Average	85.95	89.32	**91.01**	+

Kết quả phương pháp học thuộc tính của luận án chỉ đạt top 2 và top 3 so với các công trình khác. Tuy nhiên, có một số thuộc tính đạt kết quả nổi trội (bảng 4.14). Điều này minh chứng cho sự hỗ trợ của phả hệ tri thức và học đa nhiệm phát huy hiệu quả ở một số thuộc tính trên.

Tuy nhiên, kết quả học thuộc tính này được áp dụng hiệu quả cho truy vấn ảnh và cho kết quả truy vấn ảnh đạt kết quả cao được trình bày ở phần tiếp theo.

4.3.4 Thực nghiệm hệ thống truy vấn ảnh người

Kết quả thực nghiệm truy vấn ảnh cho 3 trường hợp trên như sau:

BẢNG 4.18. KẾT QUẢ THỰC NGHIỆM TRÊN TẬP PETA

Trường hợp	Top-1	Top-5	Top-10	mAP
1	**92.36**	**97.22**	**97.22**	**94.98**
2	86.00	92.80	93.20	87.66
3	40.60	48.30	49.80	40.39
Average	72.99	79.44	80.07	74.34

So sánh với các công trình điển hình trên tập PETA:

BẢNG 4.19. SO SÁNH KẾT QUẢ THỰC NGHIỆM TRÊN TẬP PETA

Công trình	Top-1	Top-5	Top-10	mAP
UPAR [109]	29.70			30.20
SAL [111]	47	66.50	74	41.50
ASMR [112]	56.50	**80.00**	**83.50**	50.20
Ours (avg)	**72.99**	79.44	80.07	**74.34**

Kết quả trên tập Pa100k:

BẢNG 4.20. KẾT QUẢ THỰC NGHIỆM TRÊN TẬP PA100K

Trường hợp	Top-1	Top-5	Top-10	mAP
1	**65.44**	**67.53**	**69.63**	**25.35**
2	62.90	64.51	66.12	15.15
3	19.40	33.60	43.60	7.22
Average	49.25	55.21	59.78	15.91

Kết quả so sánh với các công trình điển hình trên tập Pa100k:

BẢNG 4.21. SO SÁNH KẾT QUẢ THỰC NGHIỆM TRÊN TẬP PA100K

Công trình	Top-1	Top-5	Top-10	mAP
UPAR [109]	39.50			**30.50**
ASMR [112]	31.9	49.1	58.2	20.6
AIHM [113]	31.3	45.1	50.0	17.0
Ours (avg)	**49.25**	**55.21**	**59.78**	15.91

4.4 Thực nghiệm 4: hệ thống truy vấn ảnh dựa vào đa phương thức kết hợp văn bản và hình ảnh

Kết quả thực nghiệm trên tập PETA như sau:

BẢNG 4.23. KẾT QUẢ THỰC NGHIỆM TRÊN TẬP PETA

STT	Trường hợp	Top-1	Top-5	Top-10	mAP
TH1	Sử dụng đặc trưng kết hợp	26.26	58.83	76.26	30.96
TH2	Ưu tiên tìm văn bản trước	28.57	**60.79**	**77.46**	39.44
TH3	Giao kết quả	**35.59**	51.69	53.81	**42.29**
TH4	Kết hợp học thuộc tính	29.92	51.81	61.73	37.21

So sánh với các công trình khác cùng thực nghiệm trên tập dữ liệu PETA:

BẢNG 4.24. SO SÁNH KẾT QUẢ THỰC NGHIỆM TRÊN TẬP PETA

Công trình	Top-1	Top-5	Top-10	mAP
UPAR [109]	29.70			30.20
SAL [111]	47	66.50	74	41.50
ASMR [112]	56.50	**80.00**	**83.50**	50.20
Ours: IR-AL	**72.99**	79.44	80.07	**74.34**
Ours: TextIR-TH3	35.59	51.69	53.81	42.29

Bảng 4.23 và 4.24 cho thấy kết quả đề xuất của luận án chưa cao do luận án không xây dựng mô hình học. Tuy nhiên, kết quả này cao hơn công trình UPAR [109] (CVPR-2023)

Chương 5. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

5.1 Kết luận

Với nhiệm vụ "Tăng cường khả năng hiểu ảnh của hệ thống truy vấn ảnh", luận án đã hoàn thành nhiệm vụ trên. Từ hướng tiếp cận đến thiết kế các bài toán con đã đạt kết quả cao. Sau đây là các đóng góp chính của luận án:

Đóng góp thứ nhất là xây dựng thành công phả hệ tri thức thuộc tính đối tượng OAO. Luận án khảo sát các đặc điểm của mỗi loại đối tượng để đưa ra bốn loại mô tả đối tượng theo phân cấp bộ phận của đối tượng và nhóm thuộc tính của đối tượng. Đó là đối tượng chỉ có một bộ phận (như quả bóng), đối tượng có nhiều bộ phận, và nhóm mỗi bộ phận có chứa đối tượng. Dựa vào OAO, luận án thực hiện xây dựng phả hệ tri thức mô tả tri thức cho các đối tượng người HAO. Luận án thực nghiệm thành công HAO cho mô hình học thuộc tính và truy vấn ảnh người trong công trình [CT3][CT1][CT4].

Đóng góp thứ hai là xây dựng thành công mô hình học thuộc tính dựa vào phả hệ tri thức và sử dụng kết quả này hỗ trợ truy vấn ảnh [CT4].

Đóng góp thứ ba là xây dựng thành công hệ thống truy vấn ảnh dựa vào đặc trưng học sâu [CT2] và hệ thống truy vấn ảnh dựa vào phả hệ tri thức và học thuộc tính và đạt kết quả cao [CT4][CT1].

Các đóng góp trên minh chứng cho nhiệm vụ tăng cường khả năng hiểu ảnh của hệ thống truy vấn ảnh theo các hướng tiếp cận: dựa vào đặc trưng học sâu, dựa vào phả hệ tri thức và học thuộc tính để giúp hệ truy vấn ảnh hiểu ảnh tốt hơn và đáp ứng các tiêu chí "*chính xác hơn, tiện dụng và thông minh hơn*".

Luận án thể hiện tính linh hoạt và tính khả mở rộng của các giải pháp đề xuất. Công trình [CT2] là giải pháp giúp hệ thống truy vấn ảnh phát hiện đối tượng và rút trích đặc trưng học sâu cho đối tượng và truy vấn đối tượng mà không phải rút trích đặc trưng toàn ảnh, giúp tăng độ chính xác và có thể áp dụng một lớp rộng các tập dữ liệu đa dạng. Công trình [CT1][CT3][CT4] giúp hệ thống truy vấn ảnh dựa vào phả hệ tri thức thuộc tính đối tượng và học thuộc tính cho phép truy vấn ảnh từ thô đến tinh ở mức chi tiết là thuộc tính đối tượng và có thể áp dụng mở rộng cho các tập dữ liệu đặc thù như người, thời trang, cảnh vật ngoài trời và trong nhà v.v. Mô đun truy vấn ảnh dựa vào văn bản và hình ảnh cũng thể hiện tính linh hoạt trong truy vấn ảnh.

Bên cạnh kết quả đạt được, luận án vẫn còn một số hạn chế sau: Mô hình học thuộc tính vẫn chưa đạt kết quả cao nhất, chưa thực nghiệm đầy đủ các thuộc tính của đối tượng trong phả hệ tri thức thuộc tính đối tượng, chỉ thực nghiệm các thuộc tính theo bộ dữ liệu thực nghiệm. Luận án vẫn chưa khai thác tập luật để suy diễn thuộc tính giúp hệ thống phát hiện các tri thức mới mà không cần học hay huấn luyện.

5.2 Hướng phát triển

Trong tương lai, việc xây dựng hệ thống truy vấn ảnh cần phát triển theo hướng đáp ứng nhu cầu người dùng ngày càng cao.

Về dữ liệu đầu vào truy vấn: Ngoài dữ liệu đầu vào là hình ảnh và văn bản, có thể tích hợp âm thanh làm đầu vào cho câu truy vấn. Khi đó, người dùng

có thể tìm ảnh bằng giọng nói với yêu cầu truy vấn được đặc tả ở mức thuộc tính. Có thể truy vấn ảnh theo ngữ nghĩa hay ngữ cảnh.

Sử dụng các mô hình học sâu mạnh hơn để hỗ trợ hệ thống truy vấn ảnh hiệu quả hơn nữa. Khai thác mô hình ngôn ngữ ảnh (Vision-language pre-trained model) trong truy vấn ảnh.

Sử dụng phả hệ tri thức giúp tái sử dụng tri thức có sẵn là hướng đi tiềm năng giúp giảm tải cho mô hình học thuộc tính đối tượng. Tuy nhiên, các phả hệ tri thức cần có cách phát triển chung để cộng đồng cùng sử dụng chung tài nguyên. Điều này, giúp giảm chi phí thu thập và tổ chức tri thức một cách riêng lẻ. Tìm cách tăng cường khả năng hiểu ảnh khi đa dạng hóa cây phả hệ tri thức. Ngoài ra, sử dụng đồ thị tri thức để biểu diễn tri thức có sẵn là một hướng tiếp cận thú vị và tiềm năng trong truy vấn ảnh.

5.3 Các thách thức đã được giải quyết

Luận án đã giải quyết một phần nhỏ về thách thức "Lỗ hổng ngữ nghĩa". Cụ thể là luận án giúp hệ thống truy vấn tăng khả năng hiểu ảnh như sau:

- Hiểu ảnh từ mức thô (đối tượng) đến mức chi tiết (thuộc tính).
- Phân loại thuộc tính đối tượng trước và sau đó hệ truy vấn tìm kiếm đối tượng theo thuộc tính đối tượng. Điều này cũng thể hiện hệ truy vấn thông minh hơn
- Huấn luyện thuộc tính trên mỗi bộ phận của đối tượng: Cải tiến độ chính xác và giảm chi phí vì không tốn công huấn luyện thuộc tính trên những bộ phận của đối tượng không liên quan với thuộc tính đó. Từ đó giúp giảm chi phí.
- Sử dụng phả hệ tri thức giúp hiểu đối tượng, thuộc tính đối tượng và phát hiện, suy diễn mối quan hệ giữa đối tượng và đối tượng, đối tượng và thuộc tính mà không cần huấn luyện. Điều này thể hiện tính *tái sử dụng tri thức có sẵn* và giảm chi phí huấn luyện.
- Kết hợp văn bản và hình ảnh để tăng tính tiện dụng và thông minh của hệ thống

DANH MỤC CÔNG TRÌNH CÔNG BỐ CỦA TÁC GIẢ

[CT1] H. M. Nguyen, N. Q. Ly, and T. T. T. Phung, "Large-Scale Face Image Retrieval System at Attribute Level Based on Facial Attribute Ontology and Deep Neuron Network," Lect. Notes Comput. Sci., vol. 10752 LNAI, pp. 539–549, 2018.

[CT2] T. T. T. Phung, N. Q. Ly, T. T. Vo, and M. T. N. Ho, "Deep Feature Learning Network for Vehicle Retrieval," ACM International Conference Proceeding Series, pp. 18–21, 2021.

[CT3] Phung Thai Thien Trang, Fukuzawa Masayuki, &Ly Quoc Ngoc (2021), An Overview of Facial Attribute Learning," Ho Chi Minh City University of Education Journal of Science, 18(3), pp. 1859–3100, 2021.

[CT4] T. T. T. Phung, N. Q. Ly, and F. Masayuki, "A Human Retrieval System based on Human Attribute Ontology and Deep Multi-task Neural Network," Journal on Information Technologies & Communications, vol. 2, 2024. https://ictmag.vn/ict/article/view/1255