ĐẠI HỌC QUỐC GIA TP.HCM

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

NGUYỄN VINH TIỆP

TRUY VẤN HIỆU QUẢ THÔNG TIN THỊ GIÁC TỪ DỮ LIỆU LỚN ĐỂ PHÁT TRIỂN MÔI TRƯỜNG THÔNG MINH

Chuyên ngành: Khoa học Máy tính

Mã số: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

TP HỒ CHÍ MINH–Năm 2019

Công trình được hoàn thành tại:

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

Người hướng dẫn khoa học:

PGS. TS. Trần Minh Triết

PGS. TS. Dương Anh Đức

Phản biện 1: PGS. TS Nguyễn Thanh Bình

Phản biện 2: PGS. TS Trần Thị Thanh Hải

Luận án sẽ/đã được bảo vệ trước

Hội đồng chấm luận án cấp Trường tại : Đại học Công nghệ Thông tin, ĐHQG TP. Hồ Chí Minh

vào lúc 14 giờ ngày 06 tháng 08 năm 2019. Có thể tìm hiểu luận án tại:

- Thư viện Quốc gia Việt Nam
- Thư viện Trường Đại học Công nghệ Thông tin

Mục lục

Chương 1
Tổng quan

1.1 Mở đầu

Hiện nay, khối lượng dữ liệu ảnh tĩnh và video đang tăng lên ngày một nhanh chóng với rất nhiều nguồn khác nhau như: mạng xã hội, dữ liệu camera ghi nhận từ các hệ thống giám sát công cộng, gia đình. Bên cạnh đó là sự phát triển của các thiết bị di động thông minh (smartphone) và thiết bị đeo (wearable device), kho dữ liệu do người dùng tạo ra hàng ngày để ghi nhận những điều thú vị trong cuộc sống ngày càng nhiều. Điều này tất yếu dẫn đến nhu cầu phân tích để hiểu và khai thác kho dữ liệu này. Trên cơ sở đó, nhiều ứng dụng khác nhau có thể được nghiên cứu phát triển nhằm cung cấp thông tin, dịch vụ, tiện ích tốt hơn phục vụ cuộc sống con người mọi lúc và mọi nơi, tạo ra và tích hợp các tính năng thông minh vào môi trường sống hằng ngày. Đây cũng chính là mục tiêu được đặt ra ngay từ đầu và xuyên suốt quá trình thực hiện của luận án: đề xuất các giải pháp giúp khai thác thông tin hình ảnh theo các **thể thức tương tác (modality)** khác nhau một cách **tự nhiên và hiệu quả**, hướng đến xây dựng **môi trường sống tích hợp tiện ích thông minh**.

1.2 Lý do thực hiện đề tài

Từ nhu cầu thực tế đã đề cập ở phần Mở đầu và **tính đa dạng của các thể thức tương tác**, luận án này **hỗ trợ các thể thức tương tác khác nhau cho việc truy vấn thông tin thị giác**, bao gồm 4 thể thức: (i) tìm kiếm khi được người dùng cung cấp một cảnh cho trước, (ii) tìm kiếm khi có hình ảnh ví dụ của một đối tượng, (iii) tìm kiếm khi có các hình ảnh ví dụ của người và địa điểm (nhiều đối tượng), và (iv) tìm kiếm dựa trên mô tả dạng văn bản ngôn ngữ tự nhiên. Đây là một số thể thức tương tác xuất phát từ **các tình huống và trải nghiệm tự nhiên** của người dùng trong thực tế khi có nhu cầu tìm kiếm kho dữ liệu hình ảnh và video.

Khi tìm kiếm với một cảnh cho trước, các mô hình truy vấn cho loại đối tượng này thường dựa trên mô hình Bag-of-Word (BOW) với nền tảng là đặc trưng cục bộ. Trong cảnh vật thường có rất nhiều đặc trưng có tính ổn định về mặt bố cục không gian nên để **tăng cường độ chính xác** thì cần phải có bước kiểm tra ràng buộc hình

học. Không những vậy, các hệ thống còn phải **đảm bảo thời gian phản hồi hợp lý** cho người dùng.

Đối với thể thức truy vấn là ảnh ví dụ của một đối tượng cho trước, đặc biệt là các **đối tượng ít đặc trưng**, việc kiểm tra ràng buộc hình học trở nên khó khăn hơn do thiếu đặc trưng bền vững. Do đó các hệ thống thường sử dụng phương pháp kết hợp các mô hình truy vấn như BOW và thuật toán phát hiện đối tượng. Tuy nhiên, việc kết hợp này vẫn chỉ dừng lại ở mức độ đơn giản là cộng trung bình giá trị độ tương đồng của từng mô hình. Do đó cần phải có một phương pháp **kết hợp một cách hiệu quả** các điểm mạnh của từng mô hình.

Đối với thể thức truy vấn trên **nhiều đối tượng khác nhau**, cụ thể là tìm kiếm người tại một địa điểm cho trước, việc đảm bảo độ chính xác càng trở nên khó khăn hơn. Tại một thời điểm, camera ghi nhận hình ảnh thường chỉ tập trung vào một đối tượng chính nên việc đánh giá độ tương đồng bằng phương pháp kết hợp sẽ không còn hiệu quả. Do đó cần phải có một phương pháp làm **tăng độ phủ của hệ thống** ngay cả trường hợp camera không ghi nhận đầy đủ thông tin của các đối tượng cần tìm.

Nếu như ở các phần trên đề cập đến thông tin đầu vào dưới dạng hình ảnh thì trong phần này chúng tôi sử dụng dạng thông tin đầu vào khác là câu mô tả tự nhiên. Thay vì sử dụng biểu hiện về mặt thị giác để so sánh với ảnh truy vấn, hệ thống sử dụng câu mô tả đánh giá độ liên quan dựa trên các đặc trưng ngữ nghĩa (visual concept). Với mỗi ảnh hoặc đoạn video, người dùng có thể quan tâm đến rất nhiều khía cạnh ngữ nghĩa khác nhau nên cần thiết **phải có một phương pháp truy vấn khai thác rất nhiều khía cạnh ngữ nghĩa** của một tấm hình.

1.3 Mục tiêu của luận án

Mục tiêu của luận án là đề xuất một số **phương pháp truy vấn hiệu quả** với các thể thức truy vấn khác nhau từ **kho dữ liệu lớn** các ảnh tĩnh hoặc video theo những nhu cầu tìm kiếm khác nhau của người dùng. Bài toán truy vấn tổng quát được mô hình hoá bởi bốn đại lượng sau:

- $\mathcal{D}$: tập hợp các ảnh tĩnh hoặc đoạn video mà hệ thống cần truy vấn.
- $\mathcal{Q}$: thông tin truy vấn đầu vào được cung cấp bởi người sử dụng hệ thống.
- h: hàm đánh giá mức độ tương đồng giữa thông tin truy vấn với từng phần tử trong tập cơ sở dữ liệu $\mathcal{D}$.

Ba đại lượng $\mathcal{D}$, $\mathcal{Q}$ và $\mathcal{H}$ có thể tùy biến với các loại dữ liệu cần truy vấn, loại thông tin đầu vào và cách thức đánh giá tương đồng khác nhau. Tương ứng với những đại lượng này ta sẽ có một số thể thức truy vấn khác nhau. Trong luận án này chúng tôi tập trung vào bốn thực thể truy vấn chính như sau.

1.3.1 Tìm kiếm với ảnh ví dụ của cảnh vật cho trước

Thể thức đầu tiên mà luận án này đề cập đến là truy vấn với một ảnh mẫu từ tập dữ liệu ảnh tĩnh.

Đầu vào: Cho trước một tập hợp ảnh: $\mathcal{D} = \{d_1, d_2, ..., d_n\}$, với n là số lượng ảnh trong tập cơ sở dữ liệu. Truy vấn $\mathcal{Q} = (q)$ là một chuỗi bao gồm duy nhất một ảnh chụp của một cảnh vật cho trước.

Đầu ra: Hệ thống trả về một chuỗi có thứ tự các kết quả có liên quan đến ảnh truy vấn q: $\mathcal{R} = (d_{r_1}, d_{r_2}, ..., d_{r_{n_j}})$. Trong đó, n_j là số lượng phần tử phân biệt trong chuỗi kết quả trả về và $h(q, d_{r_i}) \geq h(q, d_{r_j})$ với $\forall i < j$. Hàm $h(q, d_{r_i})$ tính toán mức độ liên quan giữa ảnh mẫu q và một ảnh $d_{r_i} \in \mathcal{D}$ dựa trên **sự tương đồng về mặt thị giác (visual appearance)**.

Nói một cách khác, kết quả trả về được hiểu là ở **mức độ thực thể (instance level)**. Hình 1.1 minh họa một số tình huống có liên quan về mặt thị giác đến ảnh mẫu truy vấn. Hình a) là ảnh mẫu chụp tại một góc nhìn phía trước Nhà thờ Đức Bà. Hình b) và d) là các ảnh chụp trong cơ sở dữ liệu với một góc nhìn khác nhưng phủ phần lớn cảnh vật trên ảnh mẫu tại các thời điểm khác nhau. Hình c) và e) là các ảnh chụp của các nhà thờ có hình dáng tương tự ở Thái Lan và Hà Nội. Theo định nghĩa của truy vấn theo thực thể đối tượng của chúng tôi: Hình b) và d) là các ảnh có liên quan trong khi đó c) và e) thì không liên quan. Cũng tại Nhà thờ Đức Bà nhưng nếu chụp ở đằng sau hoặc bên trong thì cũng không được tính là có liên quan đến ảnh mẫu.

1.3.2 Tìm kiếm với ảnh ví dụ của một đối tượng

Là một sự mở rộng từ ảnh sang video mà chưa xem xét đến yếu tố về mặt thời gian, trong phần này chúng tôi định nghĩa bài toán cho tập các đoạn video (shot) với truy vấn bao gồm nhiều ảnh mẫu chụp tại các góc nhìn khác nhau của cùng một đối tượng. Cụ thể là,

Đầu vào: Cho trước một tập hợp các đoạn video: $\mathcal{D} = \{F_1, F_2, ..., F_n\}$, với n là số lượng đoạn video trong cơ sở dữ liệu. Mỗi đoạn video F_i bao gồm một tập các frame ảnh của cùng một cảnh quay. Truy vấn của dạng thể thức này được xác định bởi $\mathcal{Q} = (S, ROI)$. Trong đó, $S = \{s_1, s_2, ..., s_m\}$ và $ROI = \{b_1, b_2, ..., b_m\}$ lần lượt là

a) Ảnh truy vấn - Nhà thờ Đức Bà

b) Nhà thờ Đức Bà

c) Nhà thờ ở Thái Lan

d) Nhà thờ Đức Bà xưa

e) Nhà thờ ở Hà Nội

Hình 1.1: Ví dụ về mức độ liên quan giữa ảnh truy vấn và một số loại đối tượng.

m ảnh mẫu và đường bao phân định của **một đối tượng cần quan tâm** so với phần còn lại.

Đầu ra: Hệ thống trả về một chuỗi có thứ tự các kết quả có liên quan đến truy vấn $\mathcal{Q}$: $\mathcal{R} = (F_{r_1}, F_{r_2}, ..., F_{r_{n_j}})$. Trong đó n_j là số lượng phần tử phân biệt trong chuỗi kết quả trả về và $h(S, ROI, F_{r_i}) \geq h(S, ROI, F_{r_j})$ với $\forall i < j$. Hàm $h(S, ROI, F_{r_i})$ tính toán mức độ liên quan giữa đối tượng cần tìm (S, ROI) và một đoạn video $F_{r_i} \in \mathcal{D}$ dựa trên **sự tương đồng về mặt thị giác**. Lưu ý rằng tham số đầu vào cho hàm h lúc này là tập hợp các frame ảnh.

1.3.3 Tìm kiếm với ảnh ví dụ của người và địa điểm

Trong phần này, chúng tôi đề cập đến thể thức truy vấn dạng hỗn hợp với thông tin đầu vào bao gồm các ảnh mẫu của một người và địa điểm cho trước. Trong thực tế, khi người dùng muốn tìm lại những hình ảnh trong quá khứ của người thân gắn liền với một địa danh nào đó thì thông tin đầu vào dạng hỗn hợp này là một giải pháp phù hợp. Bài toán tìm kiếm với thể thức tương tác này được định nghĩa như sau:

Đầu vào: Cho trước một tập hợp các đoạn video: $\mathcal{D} = \{F_1, F_2, ..., F_n\}$, với n là số lượng đoạn video trong cơ sở dữ liệu. Mỗi đoạn video F_i bao gồm một tập các frame ảnh của cùng một cảnh quay. Truy vấn của dạng thể thức này được xác định bởi $\mathcal{Q} = (L, S, ROI)$. Trong đó, $L = \{l_1, l_2, ..., l_p\}$ là tập hợp bao gồm p ảnh mẫu của một địa điểm quan tâm, $S = \{s_1, s_2, ..., s_m\}$ và $ROI = \{b_1, b_2, ..., b_m\}$ lần lượt là m ảnh mẫu và đường bao phân định của **một người cần quan tâm** so với phần còn lại.

Hình 1.2 minh họa một ví dụ của loại thông tin truy vấn này. Những ảnh ở trên hàng đầu tiên là các ảnh mẫu về một quán rượu đang được quan tâm tìm kiếm. Những ảnh ở hàng thứ hai ghi nhận các góc nhìn khác nhau của cùng một người đang được quan tâm.

Hình 1.2: Ví dụ về một loại truy vấn mới bao gồm các ảnh mẫu của một vị trí (hàng phía trên) và một người (hàng phía dưới) được đánh dấu bởi đường bao màu tím.

Đầu ra: Hệ thống trả về một chuỗi có thứ tự các kết quả có liên quan đến truy vấn và được sắp xếp theo thứ tự giảm dần về mức độ liên quan. Đặt kết quả trả về của truy vấn là: $\mathcal{R} = (F_{r_1}, F_{r_2}, ..., F_{r_{n_j}})$, với n_j là số lượng phần tử phân biệt trong chuỗi kết quả trả về. Việc đánh giá mức độ liên quan giữa các ảnh mẫu (L, S, ROI) và một đoạn video F_{r_i} trong tập dữ liệu được dựa trên **sự tương đồng về mặt thị giác**.

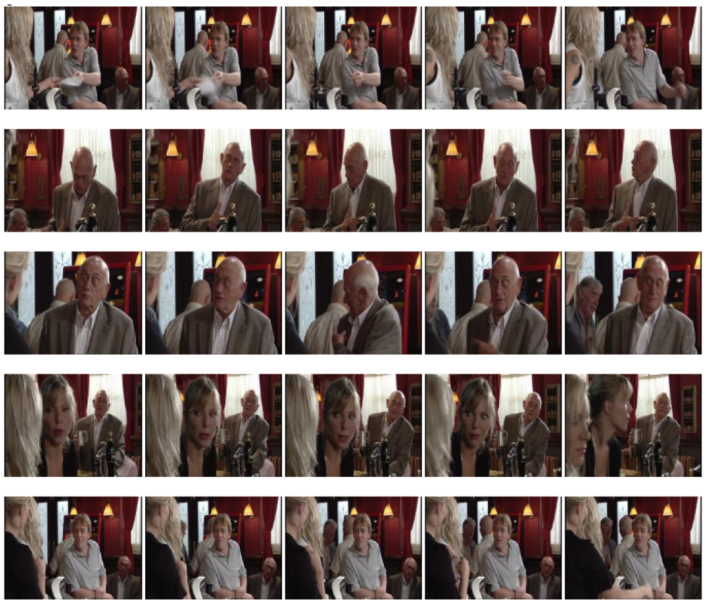
Hình 1.3 thể hiện kết quả trả về của hệ thống tìm kiếm trên loại truy vấn hỗn hợp: người tại một địa điểm cho trước. Mỗi hàng tương ứng với một đoạn video có chứa các đối tượng mô tả ở dữ liệu đầu vào. Các ảnh trên một hàng là các frame đại diện cho một đoạn video trả về. Đoạn video trả về thông thường là một phân đoạn ngắn so với tổng thể của một video có chứa đủ thông tin về mặt thị giác nhằm nhận biết được các đối tượng truy vấn.

1.3.4 Tìm kiếm dựa trên mô tả bằng ngôn ngữ tự nhiên

Trong phần này, chúng tôi đề cập đến một thể thức truy vấn không sử dụng ảnh mẫu đó chính là câu mô tả bằng ngôn ngữ tự nhiên.

Đầu vào: Cho trước một tập hợp các đoạn video: $\mathcal{D} = \{F_1, F_2, ..., F_n\}$, với n là số lượng đoạn video trong cơ sở dữ liệu. Mỗi đoạn video F_i bao gồm một tập các frame ảnh của cùng một cảnh quay. Thông tin đầu vào của dạng thể thức này được xác định bởi $\mathcal{Q} = \{c_1, c_2, ..., c_p\}$ bao gồm p từ được sử dụng để mô tả các đoạn video cần tìm.

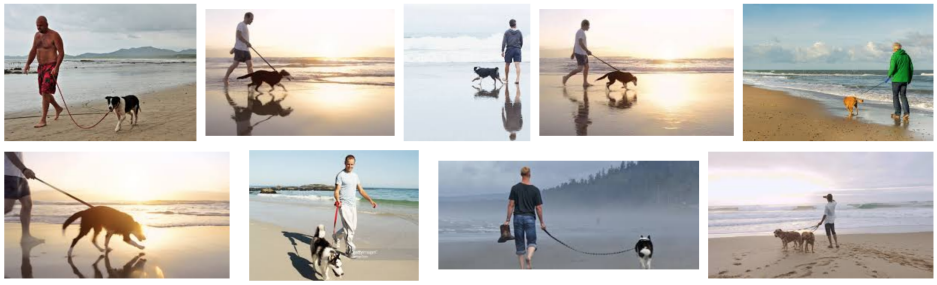
Trong luận án này, chúng tôi sử dụng các tập dataset trên tiếng Anh nên câu mô tả được viết bằng ngôn ngữ tiếng Anh. Ví dụ như:



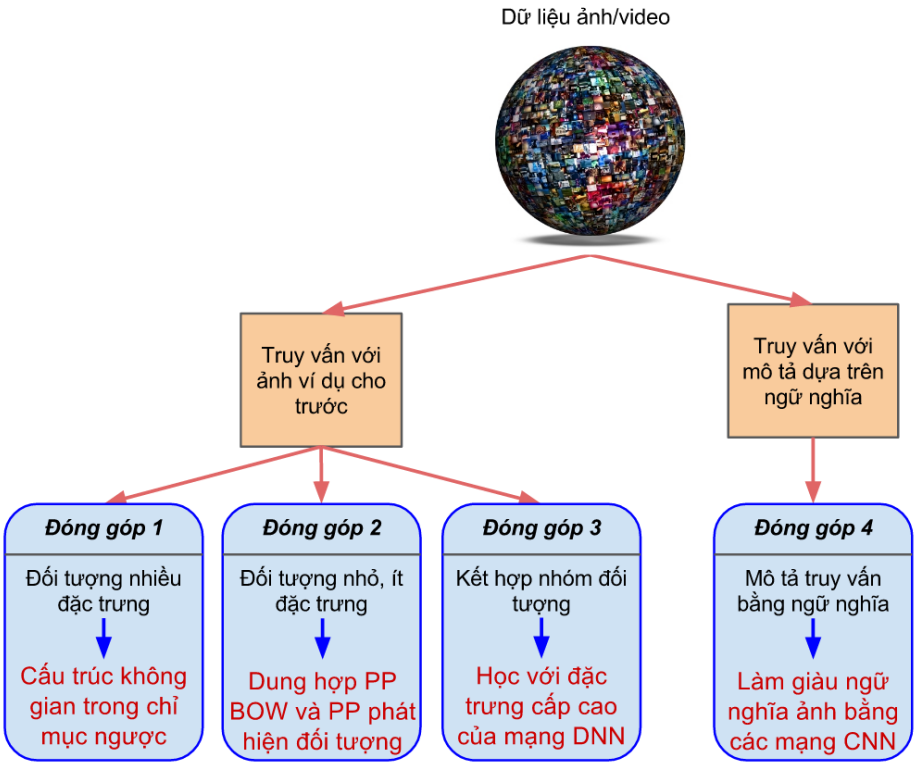
Hình 1.3: Kết quả trả về của hệ thống tìm kiếm trên thể thức hỗn hợp nhiều đối tượng: người tại một địa điểm cho trước.

"A man walking on a beach with a dog."

Đầu ra: Hệ thống trả về một chuỗi có thứ tự các kết quả có liên quan đến truy vấn và được sắp xếp theo thứ tự giảm dần về mức độ liên quan: $\mathcal{R} = (F_{r_1}, F_{r_2}, ..., F_{r_{n_j}})$. Trong đó n_j là số lượng phần tử phân biệt trong chuỗi kết quả trả về và $h(\mathcal{Q}, F_{r_i}) \geq h(\mathcal{Q}, F_{r_j})$ với $\forall i < j$. Việc đánh giá mức độ liên quan giữa câu mô tả $\mathcal{Q}$ và một đoạn video $F_{q,i}$ trong tập dữ liệu được dựa trên **sự tương đồng về mặt ngữ nghĩa hình ảnh (visual concept)**. Hình 1.4 minh hoạ kết quả trả về của hệ thống truy vấn bằng câu mô tả với nội dung có liên quan tới truy vấn.



Hình 1.4: Kết quả trả về của câu truy vấn "A man walking on a beach with a dog".



Hình 1.5: Bức tranh tổng quan của luận án.

1.4 Đóng góp chính của luận án

Luận án này tiến hành giải quyết những thách thức liên quan đến tính đa dạng của các thể thức và đối tượng truy vấn. Hình 1.5 minh họa *bức tranh tổng quan về những đóng góp của luận án.*

(i) Cấu trúc không gian trong chỉ mục ngược. Hiện nay, các hệ thống tìm kiếm đối tượng tiên tiến nhất hiện nay sử dụng ảnh mẫu đều dựa trên mô hình nền tảng là *túi từ thị giác* (Bag of Visual Word, viết tắt: BOW). Tuy nhiên, mô hình BOW dựa trên giả thuyết chính là: ***hai đối tượng ảnh tương tự nhau khi có chung một số lượng đủ lớn các vùng cục bộ (local patch) mà có thể so khớp hai chiều được với nhau***. Giả thuyết này phần nào tạo lợi thế cho việc tìm kiếm trên những đối tượng lớn có nhiều điểm đặc trưng. Ngoài ra, việc áp dụng kỹ thuật cải tiến như kiểm tra ràng buộc hình học ở bước hậu xử lý giúp tăng độ chính xác một cách đáng kể. Tuy nhiên, các kỹ thuật này thường tốn thời gian xử lý hơn và tốn bộ nhớ để lưu các thông tin liên quan đến vị trí của đặc trưng. Do đó đối với loại đối tượng lớn có nhiều đặc trưng như cảnh vật, chúng tôi đề xuất phương pháp kết hợp file chỉ mục ngược với cấu trúc kim tự tháp không gian (spatial pyramid) để vừa tăng tốc độ và giảm thời gian truy vấn. Phương pháp này được công bố trong công trình [CT1].

(ii) Dung hợp phương pháp BOW và thuật toán phát hiện đối tượng. Khi tìm kiếm với những đối tượng tương đối ít đặc trưng, ví dụ như những đối tượng nhỏ hoặc không có nhiều hoa văn, giả thuyết của mô hình BOW bị vi phạm. Do đó, chúng tôi đề xuất phương pháp kiểm tra ràng buộc mới trong đó dung hợp phương pháp BOW và phương pháp phát hiện đối tượng. Đóng góp chính của chúng tôi trong đề xuất này là khai thác hiệu quả mối quan hệ về vị trí của các *từ thị giác* (visual word) với *vị trí đề xuất đối tượng* (object instance proposal) được ước lượng bởi bộ phát hiện đối tượng. Phương pháp này được công bố trong công trình [CT3], [CT7].

(iii) Học với đặc trưng cấp cao của mạng DNN. Trong thực tế, người dùng có thể quan tâm tới rất nhiều đối tượng khác nhau cùng một lúc. Cụ thể là bài toán truy vấn trên hai loại đối tượng được quan tâm nhiều nhất là người và địa điểm. Tuy nhiên, khi ghi nhận hình ảnh, thông thường camera không tập trung vào cả hai đối tượng cùng một lúc. Do đó các đặc trưng thị giác hỗ trợ cho việc nhận biết hai loại đối tượng này sẽ không có cùng phân bố xác suất về mặt thời gian. Luận án này đề xuất phương pháp kết hợp đặc trưng học sâu với mô hình BOW và theo vết địa điểm (scene tracking) để tăng độ phủ của kết quả truy vấn. Phương pháp này được công bố trong các công trình [CT2], [CT5].

(iv) Làm giàu ngữ nghĩa ảnh bằng các mạng CNN. Đối với thể thức truy vấn sử dụng mô tả bằng từ ngữ, đây là một bài toán thú vị đang được quan tâm trong thời gian gần đây. Để giải quyết bài toán này, vấn đề mô hình hóa các khái niệm này cũng như là đánh chỉ mục để phục vụ cho bài toán truy vấn là một trong những vấn đề khó khăn cần giải quyết. Luận án này đề xuất hệ thống khai thác các đối tượng chính, các thuộc tính liên quan đến không gian, quan hệ giữa các đối tượng trong ảnh và cả dữ liệu metadata cung cấp bởi người dùng để mô tả tất cả các khía cạnh của một frame ảnh. Phương pháp này được công bố trong công trình [CT4].

Bên cạnh các thuật toán đề xuất, chúng tôi còn xây dựng các hệ thống để minh họa cho ý tưởng tương tác:

- Hệ thống khuyến nghị hỗ trợ gợi nhớ hình ảnh có liên quan dựa trên mạng xã hội dựa trên thông tin thị giác để tìm ra những hình ảnh của bạn bè hoặc chính người dùng đã từng đến một nơi nào đó trước đây [CT6].

- Hệ thống tìm kiếm ảnh đã biết trước (Known Item Search) sử dụng câu mô tả ngữ nghĩa kết hợp với các thông tin phân bố màu sắc[CT4].

Chương 2

Các công trình liên quan

2.1 Tiếp cận sử dụng đặc trưng cục bộ

2.1.1 Mô hình túi từ trong bài toán tìm kiếm đối tượng ảnh

Mô hình BOW (Bag-of-Visual-Word) được đề xuất đầu tiên bởi J. Civic và A. Zisserman [46] cho đến nay vẫn còn hiệu quả, bao gồm các bước sau:

Rút trích đặc trưng cục bộ: Có rất nhiều bộ phát hiện đặc trưng: DoG [18], Hessian-Affine [60], Harris-Laplace [59], MSER [44] cũng như bộ mô tả đặc trưng như SIFT [18], SURF [35] hay BRIEF [92]. Gần đây, R. Arandjelovíc và các đồng sự [80] đề xuất đặc trưng rootSIFT, dạng mở rộng của đặc trưng SIFT, kết hợp với độ đo khoảng cách L_2 giúp cải tiến độ chính xác mà không bộ nhớ lưu trữ.

Huấn luyện codebook: Trong nghiên cứu của Nister và Stewenius [19], sau đó là Philbin cùng các đồng sự [45] chứng minh rằng, sử dụng codebook kích thước lớn giúp tăng độ chính xác cho bài toán truy vấn đối tượng ảnh do làm giảm một cách đáng kể những cặp đặc trưng so khớp sai. Gần đây, Philbin và các cộng sự [45] đề xuất thuật toán xấp xỉ $k-$mean (approximate k-means, viết tắt AKM) sử dụng cấu trúc dữ liệu $k-d$ tree ngẫu nhiên (randomized $k-d$ tree) để xây dựng tập codebook với chi phí tính toán thấp.

Lượng tử hóa đặc trưng: Việc tăng kích thước codebook làm tăng tính phân biệt của một visual word nhưng đồng thời làm giảm tính lặp lại (repeat-ability) do các đặc trưng cục bộ chỉ hơi khác nhau nhưng được gán vào các visual word khác nhau. Để giải quyết vấn đề này, Philbin và các cộng sự [47] đề xuất sử dụng chiến lược "gán mềm" (soft assignment) trong đó mỗi vector đặc trưng được gán bởi nhiều visual word gần nhất. Sau đó, Jégou và các cộng sự [36] đề xuất phương pháp nhúng thông tin trên ba điểm neo (triangulation embedding) để biểu diễn ảnh một cách ngắn ngọn.

Xây dựng vector BOW: Việc xây dựng vector BOW theo phương pháp trọng số $tf-idf$ [46] dựa trên giả thuyết rằng trong một ảnh, vai trò của các visual word là độc lập nhau. Tuy nhiên Jégou và các đồng sự [39] đã chỉ ra rằng các visual word nếu xuất hiện trong một ảnh thì cũng sẽ có xu hướng xuất hiện lại. Tác giả đề xuất lấy căn bậc hai của thành phần số lần từ xuất hiện (term frequency - tf) để làm giảm ảnh hưởng của hiện tượng này. Phương pháp gán mềm tự nhiên cũng được sử dụng để giảm ảnh

hưởng của hiện tượng bùng nổ visual word này [3]. *Các phương pháp biến đổi trọng số tf hoặc idf trên đều được đề xuất một cách chưa tổng quát.*

So sánh hai vector BOW: Mức độ tương đồng của hai ảnh được thể hiện qua mức độ tương đồng của hai vector BOW biểu diễn. Bên cạnh các độ đo truyền thống dạng bất đối xứng, gần đây Zhu và các đồng sự [13] đã đề cập đến quan hệ bất đối xứng giữa đối tượng truy vấn và ảnh, đồng thời đề xuất một độ đo bất đối xứng để giải quyết vấn đề này.

2.1.2 Kiểm tra ràng buộc hình học

Việc kiểm tra ràng buộc hình học được thực hiện trong quá trình tính toán độ tương đồng của ảnh (*spatial ranking*) hoặc xếp hạng lại (*spatial re-ranking*).

Spatial re-ranking. Đầu tiên, mô hình BOW được thực hiện để trả về K ảnh đầu tiên có độ tương đồng với đối tượng truy vấn cao nhất. Để kiểm tra ràng buộc hình học với giả thuyết ảnh bị biến đổi affine, thuật toán RANSAC [66] được áp dụng trên một số ngẫu nhiên cặp điểm đặc trưng có cùng visual word. Zhang và các cộng sự [88] chỉ ra rằng hướng tiếp cận này giả định rằng đối tượng bị biến đổi cứng (rigid affine). Sau đó, các tác giả đề xuất sử dụng kỹ thuật đồ thị tam giác (triangulated graph) để có thể kiểm tra ràng buộc trên các đối tượng có khả năng biến dạng cao.

Spatial ranking. Trái với hướng tiếp cận Spatial re-ranking, hướng tiếp cận này ngầm kiểm tra ràng buộc hình học trong quá trình tính độ tương đồng của mô hình BOW. Do đó, ta không cần phải xác định tham số K vốn không chính xác khi thay đổi đối tượng truy vấn. Jégou và các cộng sự [38] để xuất kiểm tra ràng buộc yếu (Weak Geometric Consistency, viết tắt WGC) sử dụng cơ chế Hough Voting.

2.1.3 Tăng cường độ phủ

Kỹ thuật mở rộng truy vấn là một trong những kỹ thuật được sử dụng phổ biến để làm tăng độ phủ của hệ thống tìm kiếm văn bản [28]. Chum và các đồng sự [72] đã thử với nhiều phương pháp mở rộng đặc trưng thì nhận thấy phương pháp mở rộng truy vấn trung bình (average query expansion, viết tắt AQE) cho kết quả ổn định. Ý tưởng của phương pháp này là cộng trung bình vector $tf - idf$ của truy vấn với tập mở rộng để tạo thành một truy vấn mới. Sau đó, các tác giả tiếp tục mở rộng nghiên cứu của mình theo hướng: đảm bảo độ tin cậy của tập mở rộng và khai thác thông tin đặc trưng ngữ cảnh bên ngoài vùng đối tượng truy vấn một cách tự động.

2.1.4 Kết hợp các phương pháp

Một trong những phương pháp kết hợp đơn giản nhất khi sử dụng với nhiều đặc trưng khác nhau là cộng trung bình thứ hạng của các kết quả (rank-based average fusion). Caizhi và các cộng sự [13] kết hợp 6 loại đặc trưng tổ hợp từ 3 bộ phát hiện đặc trưng (Hessian-Affine [60], Harris Laplace [59], MSER [44]) và 2 bộ mô tả đặc trưng (rootSIFT [80], colorSIFT[53]). Kết quả cho thấy việc kết hợp ở bước hậu xếp hạng (late fusion) cho kết quả cao hơn so với tiền xếp hạng (early fusion).

Tác giả Zheng và các cộng sự nhận định rằng một đặc trưng bản thân nó đã tốt và có tính bổ sung cho các đặc trưng khác thì cũng sẽ được kỳ vọng là cải tiến độ chính xác[64]. Tuy nhiên, trong thực tế thì ta không thể biết được đặc trưng đó có tốt với một query hay không. Tác giả đề xuất giải pháp tự động nhận biết tính hiệu quả của đặc trưng thích nghi theo từng query (query adaptive) và sử dụng phương pháp học không giám sát do không biết trước thông tin về query nên không có dữ liệu gán nhãn.

Ở một hướng tiếp cận khác, Crowley và Zisserman đề xuất phương pháp kết hợp hai mô hình tìm kiếm đối tượng khác nhau: MLDS (Mid-Level Discriminative Patches)[23] với thuật toán DPM (Deformable Part Models)[77]. Trong đó MLDS là một dạng đặc trưng tương tự như biểu diễn của mô hình BOW nhưng các vùng cục bộ được chọn lựa bằng thuật toán học máy. Các tác giả đã kết hợp hai mô hình tìm kiếm đối tượng khác nhau bằng cách cộng trung bình giá trị tương đồng. Từ các kết quả nghiên cứu trên, chúng tôi nhận thấy việc kết hợp các đặc trưng, mô hình khác nhau là rất tiềm năng trong việc cải tiến độ chính xác của các hệ thống.

2.2 Tiếp cận biểu diễn ảnh sử dụng đặc trưng trích xuất từ mạng DNN

Với sự phát triển mạnh mẽ của các thuật toán máy học, một trong những kỹ thuật đột phá đang trở nên rất phổ biến gần đây đó chính là học sâu (deep learning). Lấy cảm hứng từ sự thành công của mạng CNN trong các bài toán này, chúng tôi tiến hành khảo sát và nghiên cứu một số hướng tiếp cận trong việc khai thác các kỹ thuật deep learning, đặc biệt là mạng CNN áp dụng cho bài toán truy vấn đối tượng hình ảnh.

Tác giả Donahue và các cộng sự [49] chứng minh rằng ta có thể lấy kết quả kích hoạt (activation) của những lớp kết nối đầy đủ cuối cùng làm đặc trưng biểu diễn cho bài toán nhận diện đối tượng thị giác trên một miền dữ liệu mới. Các đặc trưng thuộc lớp kết nối đầy đủ này được ký hiệu là FC (fully connected). Các đặc trưng từ kết quả

kích hoạt của những lớp không kết nối đầy đủ trước đó thường không chứa đựng đặc trưng ngữ nghĩa cao bằng các lớp cuối cùng.

Gần đây là sự xuất hiện của các thuật toán phát hiện đối tượng tiên tiến dựa trên mạng CNN như: Fast R-CNN [85] và Faster R-CNN [90]. Amaia và các cộng sự đề xuất huấn luyện lại các ảnh ví dụ với thuật toán Faster RCNN (FRCNN) như là một mô hình phát hiện đối tượng [8]. Mô hình này sử dụng ít tài nguyên tính toán hơn phương pháp BLCF do thời gian huấn luyện ít hơn. Luận án này khai thác sức mạnh của các mô hình phát hiện đối tượng và mô hình truy vấn đối tượng với đặc trưng cục bộ được trình bày trong công trình [CT3], [CT7].

2.3 Tiếp cận biểu diễn ảnh sử dụng ngữ nghĩa

Bài toán biểu diễn ảnh bằng ngữ nghĩa được cộng đồng khoa học đặt ra ngay từ thời kỳ đầu của ngành thị giác máy tính. Mục tiêu mà bài toán đặt ra đó là làm cho máy tính có thể hiểu và diễn đạt được một bức ảnh dưới khía cạnh ngữ nghĩa.

Đối với hướng tiếp cận biểu diễn ngữ nghĩa bằng văn bản, nhiều hướng tiếp cận đã lấy cảm hứng từ các quả nghiên cứu thành công trong việc sử dụng mạng *neural network phản hồi* (Recurrent Neural Network, viết tắt là RNN) trong việc huấn luyện đối sánh chuỗi của bài toán dịch máy. Có thể kể đến như các nghiên cứu của Cho và các đồng sự [62], Bahdanau và các đồng sự [22], Sutskever và các đồng sự [43]. Một trong những lý do chính mà bài toán gán phụ đề cho ảnh khá phù hợp với mô hình mã hóa - giải mã (encoder-decoder) trong mô hình dịch máy của [62] là vì sự tương đồng trong việc diễn dịch một ảnh sang câu văn.

Những hướng tiếp cận mới nhất hiện nay đều sử dụng mạng RNN được đề xuất sử dụng trước đó bởi Werbos [76], Hochreiter và Schmidhuber [89] làm cốt lõi. Tuy nhiên các hướng tiếp cận này chỉ xem xét các đối tượng trong ảnh tại một thời điểm, Vinyals và các cộng sự [74], Donahue và các cộng sự [49] sử dụng các mạng Long short-term memory RNN (LSTM RNN) trong mô hình biểu diễn và phát sinh mô tả ngữ nghĩa trong cả ảnh lẫn video. Luận án này tập trung theo hướng tiếp cận này để làm giàu ngữ nghĩa cho ảnh và xây dựng hệ thống truy vấn với thể thức sử dụng câu mô tả.

Chương 3
Kết hợp Spatial Pyramid và cấu trúc chỉ mục ngược cho bài toán tìm kiếm cảnh vật

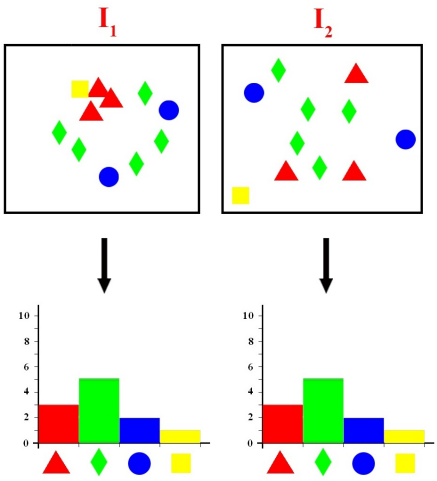
3.1 Mở đầu

3.1.1 Nhu cầu thực tế

Hiện nay, các tổ chức cũng như nhóm nghiên cứu trong và ngoài nước đang quan tâm đến việc xây dựng các ứng dụng thông minh nhằm hướng đến thành phố thông minh (smart city). Dữ liệu do người dùng tạo ra thông qua các thiết bị di động đang ngày càng trở nên phổ biến, đặc biệt là dữ liệu hình ảnh. Người dùng chụp lại những cảnh vật mà họ ấn tượng để sau này có thể tìm kiếm và xem lại về sau. Tuy nhiên, khi người dùng muốn tìm kiếm một đối tượng bằng từ khoá nhưng không biết trước tên gọi hoặc dữ liệu ảnh không được gán nhãn thì thể thức truy vấn với ảnh mẫu là một giải pháp phù hợp. Đối tượng truy vấn đầu tiên mà chương này hướng đến là cảnh vật, bao gồm một hoặc nhiều đối tượng khác nhau có cấu trúc không gian cố định theo thời gian.

3.1.2 Hướng tiếp cận của luận án

Có rất nhiều nghiên cứu đã được đề xuất cho bài toán truy vấn ảnh trên cảnh vật. Đa số trong số đó là dựa trên mô hình Bag-of-Word (BOW) được đề xuất bởi Sivic và Zisserman năm 2003 [46]. Tuy nhiên, giới hạn chính của mô hình BOW truyền thống và cấu trúc chỉ mục ngược là đã loại bỏ nguồn thông tin quan trọng giúp phân biệt chính xác ảnh truy vấn và ảnh trong cơ sở dữ liệu. Hình 3.1 minh họa một ví dụ điển hình của hiện tượng này. Hướng tiếp cận chính của chương này là tận dụng thông tin về không gian của visual word của mỗi ảnh để làm tăng độ chính xác trong khi vẫn đảm bảo được thời gian truy vấn ngắn. Ý tưởng chính là xem xét đến độ phân bố của các đặc trưng trên từng vùng của ảnh thông qua kết hợp cấu trúc chỉ mục ngược để xác định nhanh chóng những vùng có liên quan.

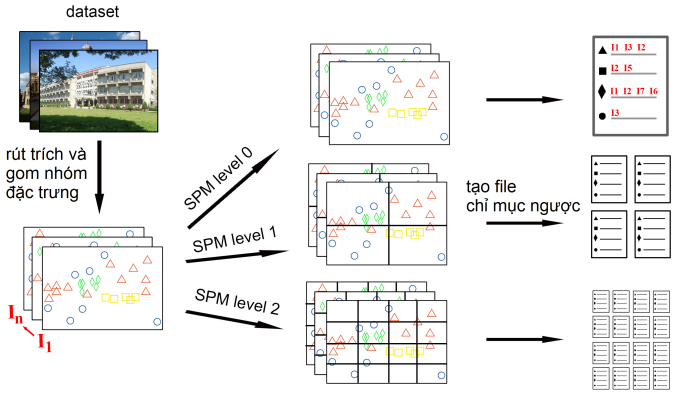


Hình 3.1: I_1và I_2 giống nhau về đặc trưng histogram nhưng khác nhau về thị giác.

3.2 Kết hợp cấu trúc không gian với chỉ mục ngược

3.2.1 Tích hợp thông tin không gian vào cấu trúc chỉ mục ngược sử dụng Spatial Pyramid

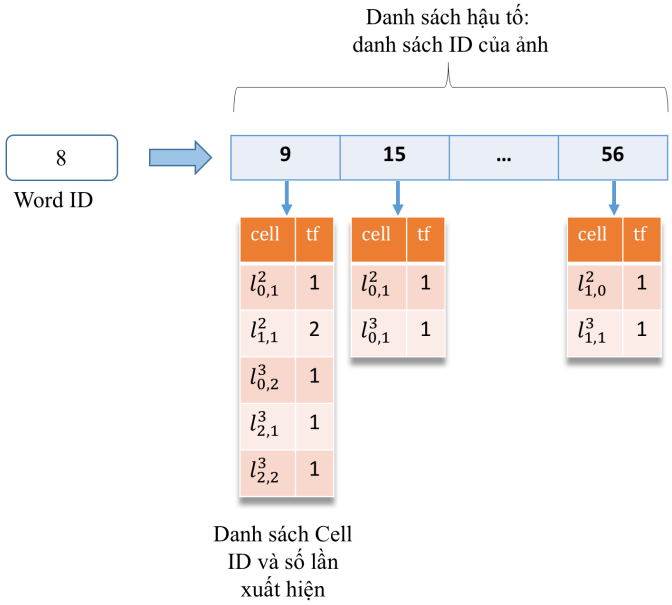
Ý tưởng chính là của phương pháp đề xuất là chia ảnh thành nhiều vùng của dạng lưới kim tự tháp (hay còn gọi là Spatial Pyramid). Một lưới tại mức l của kim tự tháp chia ảnh thành $2^l \times 2^l$ ô có kích thước giống nhau. Histogram của các visual word thuộc các ô được xây dựng và nối lại với nhau với các trọng số để tạo thành vector đặc trưng cuối cùng đại diện cho một ảnh. Hình 3.2 minh họa quá trình xử lý Offline của hệ thống.



Hình 3.2: Tổng quan của phương pháp đề xuất.

Tại bước xử lý offline, chúng tôi sử dụng Spatial Pyramid để chia ảnh thành các

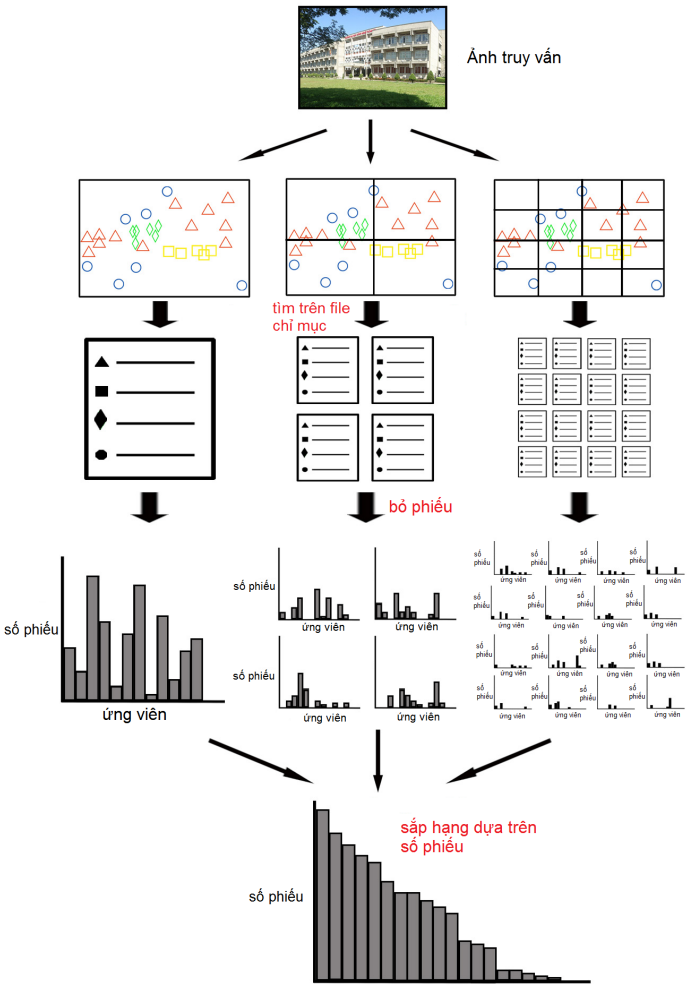
ô dựa trên mức cho trước và gom tất cả các word nằm trong mỗi ô. Bước tiếp theo, tập hợp các word trong mỗi ô của ảnh được sử dụng để phát sinh cấu trúc chỉ mục ngược. Cho trước giá trị mức tối đa của Spatial Pyramid là L. Số lượng file chỉ mục ngược là $\frac{1}{3}(4^{L+1}-1)$ và mỗi mức sẽ có $2^l \times 2^l$ file với $0 \leq l \leq L$. Trong quá trình tạo chỉ mục ngược, thông tin về không gian của các visual word được lưu kèm theo số lần xuất hiện tại một ô thuộc không gian Spatial Pyramid.



Hình 3.3: Cấu trúc chỉ mục ngược có tích hợp thông tin cấu trúc không gian.

Tại bước xử lý online, ảnh truy vấn cũng được thực hiện một cách tương tự như trên ảnh thuộc cơ sở dữ liệu bao gồm: rút trích đặc trưng và lượng tử hóa đặc trưng ảnh. Dựa trên vị trí của mỗi visual word, chúng tôi gán vào các ô tương ứng trong lưới với các mức khác nhau của cấu trúc Spatial Pyramid. Mỗi visual word được sử dụng để truy xuất trực tiếp vào cấu trúc chỉ mục ngược đề xuất để đồng thời lấy thông tin posting list và xếp hạng lại các ảnh ứng viên trên các ô có liên quan. Chúng tôi sử dụng cơ chế voting để xếp hạng cho các ảnh, trong đó giá trị độ tương đồng sử dụng để xếp hạng được tính toán trong quá trình truy xuất các cấu trúc chỉ mục ngược như mô tả ở Hình 3.4.

Thuật toán 1 và 2 lần lượt trình bày các bước chính của hai thuật toán xây dựng cấu trúc chỉ mục tích hợp thông tin không gian và thuật toán tìm kiếm với cấu trúc chỉ mục ngược này.



Hình 3.4: Chi tiết quá trình truy vấn ảnh.

3.3 Kết quả thực nghiệm

3.3.1 Dataset và độ đo đánh giá

Dataset: Chúng tôi sử dụng 3 dataset chuẩn để đánh giá bao gồm: Oxford 5K, Oxford 105K và Paris 6K. Các dataset Oxford 5K và Paris 6K lần lượt bao gồm 5062 ảnh và 6412 ảnh có độ phân giải cao. Oxford 105K là tập mở rộng của Oxford 5K với khoảng 100,000 ảnh làm nhiễu được download tự động từ trang Flickr.

Độ đo đánh giá: Để đánh giá độ chính xác của hệ thống truy vấn, chúng tôi sử dụng độ đo Độ chính xác trung bình (Mean Average Precision – MAP). Bên cạnh đó, chúng tôi còn tiến hành đo lường tốc độ thực thi truy vấn trung bình cho mỗi truy vấn. Do tất cả các hướng tiếp cận đều dựa trên mô hình BOW để biểu diễn nên chúng tôi không quan tâm đến phần thời gian của quá trình rút trích đặc trưng. Do đó, chúng

Thuật toán 1: Đánh chỉ mục cho một ảnh đầu vào.

InsertBowToSPIndex $(B, id, Levels, SPIndex)$

Đầu vào: $B = (x_i, y_i, w_i)_{i=1..n}$: tập hợp các visual word của một ảnh
n: số lượng visual word của một ảnh
x_i, y_i: thông tin vị trí của visual word thứ i
w_i: định danh của visual word thứ i
id: định danh của ảnh/shot
$Levels = 1, 2, ..., L$: tập hợp các mức phân vùng ảnh
$SPIndex$: cấu trúc chỉ mục ngược với thông tin không gian

$base\_fid \leftarrow 0$
foreach $l \in Levels$ **do**
    // Tính kích thước cell
    $wcell \leftarrow W/l$
    $hcell \leftarrow H/l$
    foreach $(x_i, y_i, w_i) \in B$ **do**
        // Tính ID của ô đang chứa đặc trưng
        $q_{x_i} \leftarrow \lfloor x_i/wcell \rfloor$
        $q_{y_i} \leftarrow \lfloor y_i/hcell \rfloor$
        $fid \leftarrow base\_fid + q_{x_i} * l + q_{y_i}$
        $SPIndex[w_i]$.push($(id,\ fid)$)
    base_fid ← base_fid + l^2

tôi tiến hành đo và lưu thời gian thực hiện kể từ sau khi rút trích đặc trưng đến lúc kết thúc quá trình xếp hạng kết quả trả về.

3.3.2 Cấu hình các thí nghiệm

Cấu hình của các thí nghiệm được mô tả như sau:

- Baseline 1: Sử dụng mô hình BOW với cấu trúc chỉ mục ngược để xếp hạng kết quả trả về dựa trên cơ chế voting.
- Baseline 2: Sử dụng mô hình BOW kết hợp với cấu trúc chỉ mục ngược và xếp hạng lại sử dụng công thức tính độ đo khoảng cách giữa ảnh truy vấn và ảnh ứng viên. Mỗi ảnh được biểu diễn sử dụng mô hình Spatial Pyramid. Trong baseline này, chúng tôi sử dụng độ đo bất đối xứng [13] để tính toán.
- II+SPM: Sử dụng mô hình BOW với cấu trúc chỉ mục ngược tích hợp thêm thông tin không gian trong quá trình đánh chỉ mục và xếp hạng lại.

Thuật toán 2: Thuật toán truy vấn trên cấu trúc chỉ mục không gian.

result = **QueryWithSPIndex** $(B, Levels, SPIndex)$

Đầu vào: $B = (x_i, y_i, w_i)_{i=1..n}$: tập hợp các visual word của truy vấn
n: số lượng visual word của truy vấn
x_i, y_i: thông tin vị trí của một visual word thứ i
w_i: định danh của visual word thứ i
$Levels = 1, 2, ..., L$: tập hợp các mức phân vùng ảnh
$SPIndex$: cấu trúc chỉ mục ngược với thông tin không gian
w_{bg}: trọng số của các visual word thuộc vùng nền (background) khi xây dựng đặc trưng Bag-of-word

Đầu ra : $result$: danh sách các ảnh/đoạn video và độ tương đồng được sắp giảm dần theo mức độ liên quan

```
result = []
scores = []
foreach l ∈ Levels do
    wcell ← W/l
    hcell ← H/l
    foreach (x_i, y_i, w_i) ∈ B do
        q_{x_i} ← ⌊x_i/wcell⌋
        q_{y_i} ← ⌊y_i/hcell⌋
        posting ← SPIndex[w_i, q_{x_i}, q_{y_i}]
        // Lọc lại các visual word thuộc
        // Tính toán giá trị tương đồng trên từng level
        scores[q_{x_i}, q_{y_i}] ← ComputeScore(B, posting, w_bg)
f_score ← mean(scores) //Tổng hợp độ tương đồng trên tất cả level
result ← sort(f_score)
return result
```

- II+SPM*: Kết hợp II+SPM với ô trung tâm để đánh chỉ mục và xếp hạng. Điểm trái trên và phải dưới của ô trung tâm lần lượt đặt tại các vị trí có tọa độ $(\frac{w}{4}, \frac{h}{4})$ và $(\frac{3w}{4}, \frac{3h}{4})$, trong đó w và h lần lượt là chiều rộng và chiều dài của ảnh.

3.3.3 Kết quả thực nghiệm

Tính hiệu quả của việc kết hợp Spatial Pyramid và cấu trúc chỉ mục ngược: Bảng 3.1 trình bày chi tiết các kết quả của bốn hệ thống được khi tiến hành thí nghiệm trên ba dataset: Oxford 5K, Oxford 105K và Paris 6K. Bảng 3.2 ghi lại chi tiết thời gian thực thi của tất cả các hệ thống trên các dataset. Các kết quả thí nghiệm cho thấy, phương pháp đề xuất II+SPM cân bằng được giữa độ chính xác với độ đo MAP và thời gian truy vấn tính từ lúc nhận ảnh mẫu đầu vào đến lúc trả kết quả.

Ngoài ra, dựa vào các bảng kết quả ta có thể thấy, đề xuất thứ hai II+SPM* cho kết quả cao hơn đáng kể so với các cấu hình còn lại trên cả ba dataset và thời gian truy vấn thấp hơn rất nhiều so với Baseline 2 và xấp xỉ so với Baseline 1 và II+SPM.

Bảng 3.1: Độ chính xác của các phương pháp trên các tập dữ liệu.

	Baseline 1	Baseline 2	II+SPM	II+SPM*
Oxford 5K	0.6258	0.6333	0.6318	**0.6564**
Oxford 105K	0.5176	0.5523	0.5494	**0.5944**
Paris 6K	0.6133	0.6273	0.6234	**0.6604**

Bảng 3.2: Thời gian truy vấn của các phương pháp trên các tập dữ liệu.

	Baseline 1	Baseline 2	II+SPM	II+SPM*
Oxford 5K	0.10	11.53	0.15	0.19
Oxford 105K	3.13	21.02	4.39	4.42
Paris 6K	0.17	13.29	0.29	0.32

3.4 Kết luận

Trong chương này, chúng tôi trình bày hướng tiếp cận tích hợp thông tin không gian của đặc trưng vào cấu trúc chỉ mục ngược để cải tiến độ chính xác trong khi vẫn giữ được thời gian truy vấn ngắn. Ngoài ra, nghiên cứu sinh xây dựng hệ thống khuyến nghị gợi nhớ hình ảnh sử dụng đặc trưng thị giác cho người dùng mạng xã hội. Kết quả cho thấy người dùng đều có trải nghiệm thú vị và tích cực khi tham gia sử dụng ứng dụng đề xuất này. Hệ thống này được mô tả chi tiết trong công trình [CT6].

Chương 4
Dung hợp mô hình BOW và thuật toán phát hiện đối tượng cho bài toán tìm kiếm đối tượng ít đặc trưng

4.1 Mở đầu

4.1.1 Nhu cầu thực tế

Các chương trước đề cập đến bài toán truy vấn ảnh mẫu cảnh vật bao gồm những đối tượng lớn với nhiều đặc trưng thị giác. Tuy nhiên trong nhiều tình huống thực tế, đối tượng mà người dùng quan tâm chỉ là một đối tượng nhỏ trong một tấm ảnh như: logo của một mặt hàng, sản phẩm mẫu, vật dụng cá nhân. Việc xây dựng các hệ thống truy vấn trên các đối tượng đơn giúp hệ thống truy vấn không bị nhầm lẫn với các đối tượng khác có trong ảnh, nâng cao hiệu quả tìm kiếm. Giải quyết bài toán truy vấn một đối tượng trong ảnh có thể tạo ra được các ứng dụng tiềm năng như: thương mại điện tử, quản lý thương hiệu trên mạng xã hội, tìm kiếm vật bị thất lạc và giám sát hệ thống camera.

4.1.2 Hướng tiếp cận của luận án

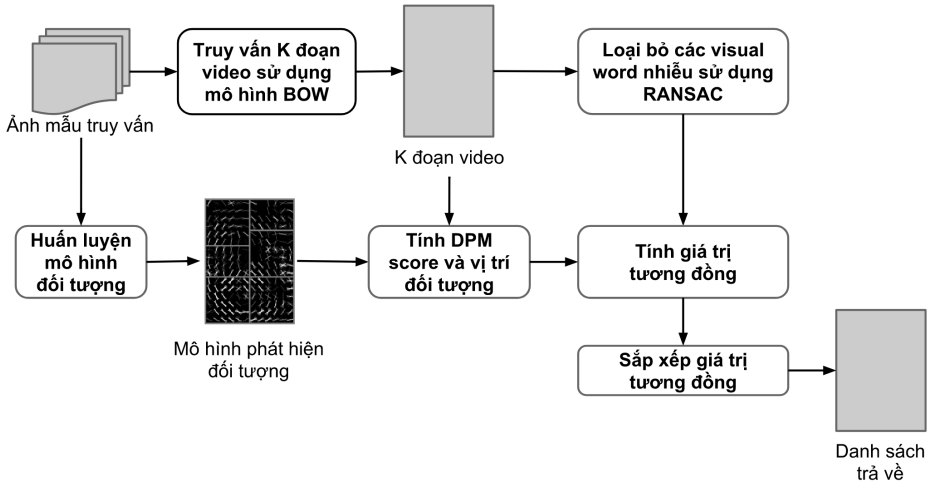
Hướng tiếp cận chính của chương này bao gồm: (i) đề xuất phương pháp tính trọng số kết hợp sử dụng mạng neural network để khai thác đặc điểm bên ngoài của đối tượng truy vấn, (ii) giới thiệu phương pháp kết hợp mới khai thác vị trí tương đối của các cặp visual word tương đồng với đường bao của đối tượng xác định bởi thuật toán phát hiện đối tượng. Chúng tôi chia độ tin cậy của các visual word thành 3 loại. Ứng với mỗi loại sẽ có một hàm trọng số đóng góp vào hàm tổng hợp độ tin cậy cuối cùng. Không giống như hàm trọng số *idf* được tính toán trên toàn bộ tập dữ liệu, hàm trong số được chúng tôi đề xuất có khả năng thích nghi theo từng đặc điểm của đối tượng truy vấn cụ thể.

4.2 Dữ liệu thử nghiệm và phương pháp đánh giá

Trong chương này, chúng tôi tiến hành thực nghiệm trên tập dữ liệu TRECVID Instance Search (TRECVID INS). Kích thước dữ liệu khoảng 300GB với thời lượng lên đến 464 giờ, gồm 244 file video được trích từ kênh BBC EastEnders. Độ đo được sử dụng để đánh giá độ chính xác của hệ thống là MAP.

4.3 Hệ thống tìm kiếm đối tượng

4.3.1 Tổng quan hệ thống



Hình 4.1: Hệ thống tìm kiếm đối tượng trong kho dữ liệu video lớn.

Để biểu diễn đơn giản, chúng tôi chỉ xem xét đến các ảnh mẫu của đối tượng truy vấn và các frame chính của một đoạn video. Việc tính giá trị tương đồng cho các đoạn video khác được thực hiện hoàn toàn tương tự. Đặt $Q_k, F_j \in \mathbb{R}^L$ là vector BOW biểu diễn của ảnh mẫu thứ k của đối tượng và frame thứ j của đoạn video. L là kích thước của codebook. Để xây dựng cấu trúc chỉ mục với biểu diễn ngắn gọn, chúng tôi sử dụng phương pháp tổng hợp trung bình (average pooling):

$$F = \frac{1}{n} \sum_{j=1}^{n} F_j \tag{4.1}$$

trong đó, n là số lượng frame chính trong đoạn video. Giá trị tương đồng giữa đối

tượng truy vấn và đoạn video được tính bằng công thức sau đây:

$$S_{BOW} = \frac{1}{n'} \sum_{k=1}^{n'} asym(Q_k, F_j) \tag{4.2}$$

trong đó, n' là số lượng ảnh mẫu của đối tượng truy vấn và $asym$ là một hàm tính độ tương đồng bất đối xứng đề xuất bởi [13]. Top K đoạn video trả về dựa trên giá trị tương đồng trên (S_{BOW}) sau đó được sử dụng cho bước xếp hạng lại sau này.

Hình 4.1 thể hiện các bước của hệ thống tìm kiếm đối tượng được sử dụng trong luận án này. Module màu đỏ là mô hình nền tảng với những tham số cấu hình đã được mô tả như ở phần trên. Các module còn lại được đề xuất để giải quyết vấn đề của kiểm tra ràng buộc hình học sẽ được đề cập ở những phần sau.

4.3.2 Xác định vị trí đối tượng truy vấn với thuật toán phát hiện đối tượng

Để đánh giá mức độ tin cậy của kết quả trả về bởi mô hình BOW, chúng tôi đề xuất sử dụng kết hợp với hướng tiếp cận kiểm tra ràng buộc từ trên xuống. Minh họa cho ý tưởng này, chúng tôi sử dụng một trong những thuật toán phát hiện đối tượng cho kết quả tốt trong những năm gần đây là thuật toán Faster-RCNN (FRCNN). Thuật toán này hoàn toàn có thể thay thế bởi những thuật toán khác mà không làm thay đổi cấu trúc của chương trình.

Đặt M là bộ phát hiện đối tượng được huấn luyện từ tập ảnh mẫu truy vấn (positive) và tập ảnh ngẫu nhiên từ trang web Google Images (negative). Giá trị tương đồng giữa đối tượng truy vấn và đoạn video trong tập dữ liệu được tính như sau:

$$S_{OD} = \max_j score(M, I_j) \tag{4.3}$$

trong đó, I_j là frame chính thứ j của đoạn video. Để kết hợp mô hình BOW và thuật toán FRCNN, một trong những ý tưởng đơn giản nhất là cộng trung bình các giá trị tương đồng chuẩn hóa như công thức sau:

$$S_{new} = w_1 \cdot S^*_{BOW} + w_2 \cdot S^*_{FRCNN} \tag{4.4}$$

trong đó, $w_1 = w_2 = 1$, S^*_{BOW} và S^*_{FRCNN} là các giá trị chuẩn hóa của S_{BOW} và S_{FRCNN}

trên top K danh sách xếp hạng trả về của mỗi thuật toán, sử dụng hàm z-score:

$$s^* = norm(s) = \frac{s - \mu}{\sigma} \tag{4.5}$$

trong đó, μ và σ lần lượt là các giá trị trung bình và độ lệch chuẩn của top K các giá trị tương đồng.

4.4 Dung hợp mô hình BOW với thuật toán phát hiện đối tượng sử dụng mạng neural network

Thuật toán 3: Thuật toán học mạng xác định hệ số kết hợp dựa trên đặc điểm truy vấn.

model = **LearnParameterWithNeuralNetwork** $(query\_feat, gt\_param)$

Đầu vào: $query\_feat$: ma trận dữ liệu đặc điểm của tập đối tượng truy vấn
gt_param: mảng chứa dữ liệu tham số kết hợp tối ưu

Đầu ra : $model^*$: mạng neural network đã được huấn luyện

Bước 1: Khởi tạo mạng neural network

$model \leftarrow$ GenerateNetwork()

Bước 2: Rút trích đặc trưng truy vấn

Với mỗi truy vấn, rút trích đặc trưng thể hiện tính chất truy vấn:
$N_{key}, mean(N_{shared}), max(N_{shared}), \frac{S_{object}}{S_{image}}, mean(\frac{N_{shared}}{N_{key}}), max(\frac{N_{shared}}{N_{key}})$

Kết thúc bước này sẽ là ma trận thể hiện tính chất các truy vấn: $F \in \mathbb{R}^{d \times n}$, với d là số chiều của tính chất ảnh và n là số lượng mẫu

Bước 3: Xác định tham số kết hợp tối ưu

Với mỗi truy vấn, xác định độ tương đồng dựa trên hai mô hình BOW và thuật toán phát hiện đối tượng

Kết thúc bước này là vector các hệ số kết hợp tối ưu: $p^* \in \mathbb{R}^n$, với n là số lượng mẫu

Bước 4: Huấn luyện mạng neural network

$model^* =$ train($model$, F, p^*)

return $model^*$

Ý tưởng chính của chúng tôi là, từ tập hợp các bộ trọng số tối ưu cho rất nhiều loại đối tượng truy vấn, chúng tôi sẽ huấn luyện mô hình ước lượng giá trị trọng số tốt từ các đặc điểm của một đối tượng truy vấn bất kỳ. Trong phần này, chúng tôi sử dụng mạng neural network để hiện thực hóa ý tưởng kết hợp linh động các giá trị tương đồng của mô hình BOW và thuật toán phát hiện đối tượng.

Thuật toán 4: Thuật toán hậu xử lý dựa trên trọng số thích nghi của đối tượng truy vấn.

result = **AdaptivePostProcessing** $(Query\_wei, S^*_{BOW}, S^*_{OD})$

Đầu vào: wei: trọng số thích nghi của truy vấn
S^*_{BOW}: độ tương đồng đã chuẩn hóa sử dụng mô hình BOW
S^*_{OD}: độ tương đồng đã chuẩn hóa sử dụng thuật toán phát hiện đối tượng

Đầu ra : $result$: danh sách các ảnh/đoạn video và độ tương đồng được sắp giảm dần theo mức độ liên quan

$result = []$
$scores = []$
for $i \in [1, 2, ..., lenght(S^*_{BOW})]$ **do**
    $scores[i] = wei \cdot S^*_{BOW}[i] + (1 - wei) \cdot S^*_{OD}[i]$
$result \leftarrow$ Sort($scores$) // Sắp xếp độ tương đồng giảm dần
return $result$

Chúng tôi đề xuất 6 tính chất được chia thành 2 nhóm để học tất cả các ảnh ví dụ của đối tượng truy vấn: kích thước đối tượng và mức độ phức tạp của hoa văn đặc trưng (chi tiết trong Bảng 4.2). Chúng tôi sử dụng một mạng neural network với ba lớp: lớp input, lớp ẩn, lớp output. Số lượng neuron trong lớp input bằng với số lượng thuộc tính rút trích từ m ảnh mẫu của đối tượng truy vấn.

Đề huấn luyện mạng neural network, chúng tôi tạo một tập các mẫu trọng số tốt tương ứng cho các đối tượng truy vấn của Trecvid INS2013 và INS2014. Sau đó chúng tôi đánh giá MAP của tất cả các truy vấn sử dụng các trọng số kết hợp được ước lượng từ mạng neural network cho cấu hình tốt nhất với thông tin đầu vào cần thiết. Thuật toán 3 trình bày vắn tắt các bước thực hiện của thuật toán học tham số kết hợp tối ưu sử dụng mạng Neural Network. Thuật toán 4 mô tả bước hậu xử lý kết hợp giá trị độ tương đồng của phương pháp dựa trên BOW và thuật toán phát hiện đối tượng dựa trên hệ số kết hợp tối ưu.

4.4.1 Thí nghiệm và kết quả

Chúng tôi tiến hành hai thí nghiệm để đánh giá các ưu điểm của hệ thống tìm kiếm đối tượng đề xuất với các mô hình cơ bản BOW và thuật toán phát hiện đối tượng sử dụng phương pháp sử dụng hệ số thích nghi. Bảng 4.1 còn cho thấy phương pháp cộng trung bình cho độ đo chính xác MAP cao hơn so với hai hệ thống cơ sở. Ngoài ra, Bảng 4.3 còn cho thấy phương pháp của chúng tôi tốt hơn các phương pháp state-of-the-art.

Bảng 4.1: So sánh giữa các phương pháp kết hợp với các tham số cứng.

Phương pháp	**MAP**	
	INS2013	**INS2014**
Phát hiện đối tượng ($\alpha = 0$)	19.55	21.23
Mô hình BOW ($\alpha = 1$)	27.91	25.01
Kết hợp trung bình ($\alpha = 0.5$)	32.18	28.21

Bảng 4.2: So sánh ảnh hưởng của việc chọn các đặc trưng đầu vào cho mạng neural network lên kết quả tìm kiếm.

Số đặc trưng	Mean $\frac{S_{object}}{S_{image}}$	Mean N_{shared}	Max N_{shared}	Mean $\frac{N_{shared}}{N_{key}}$	Max $\frac{N_{shared}}{N_{key}}$	N_{key}	**MAP**	
							2013	**2014**
6	x	x	x	x	x	x	32.25	28.51
2		x	x				32.45	28.87
2	x	x					32.78	28.42
3	x			x	x		32.84	28.68
4	x	x	x	x			32.87	28.42
2	x		x				32.92	28.47
3	**x**	**x**	**x**				**33.07**	**28.67**

Bảng 4.3: So sánh với các phương pháp trên hai tập INS2013 và INS2014.

Phương pháp	**MAP**	
	INS2013	**INS2014**
Average fusion ($\alpha = 0.5$)	32.18	28.21
Multi-features[21]	31.33	28.77
HE+WGC[38]	26.51	24.34
TC[95]	20.50	N/A
PP đề xuất (Adaptive fusion)	**33.07**	**28.67**

4.5 Dung hợp mô hình BOW với thuật toán phát hiện đối tượng sử dụng quan hệ vị trí đặc trưng

Tiếp nối phương pháp trên, trong phần này chúng tôi đề xuất một phương pháp kết hợp tận dụng một cách hiệu quả các thông tin hữu ích là các đặc trưng chung của mô hình BOW và thông tin vị trí xác định bởi thuật toán phát hiện đối tượng. Hình 4.2 mô tả các trường hợp có thể xảy ra đối với đặc trưng chung của mô hình BOW và vị trí xác định bởi thuật toán phát hiện đối tượng. Ô màu trắng trên ảnh truy vấn đánh dấu vùng có chứa đối tượng truy vấn khi đưa vào hệ thống tìm kiếm. Ô màu trắng trên ảnh database là vùng vị trí mà thuật toán phát hiện được. Các cặp điểm so khớp này được chúng tôi chia ra làm loại như sau:

Loại thứ nhất: các cặp điểm so khớp sai biểu diễn bởi mũi tên màu đỏ. Các cặp

điểm này sẽ được loại bỏ bởi thuật toán RANSAC, còn lại là các cặp điểm so khớp đúng bao gồm ba loại tiếp theo.

Loại thứ hai: các cặp điểm có *tính phân biệt cao* (discriminative) được biểu diễn bằng những đường mũi tên màu xanh lá và có hàm trọng số đồng biến theo số lượng điểm $f_1(N_d)$.

Loại thứ ba: các cặp điểm *có liên quan yếu* (weakly relevant) được biểu diễn bởi những đường mũi tên màu xanh dương và có hàm trọng số đồng biến theo số lượng điểm $f_2(N_w)$. Tuy nhiên, hàm số này sẽ không tăng nhanh bằng hàm f_1.

Loại thứ tư: các cặp điểm mang thông tin ngữ cảnh (context information) được biểu diễn bằng những đường mũi tên màu đen và có hàm trọng số được sử dụng $f_3(N_c)$ không tăng nhanh bằng hàm số f_1 và f_2. Ta có công thức tính score như sau:

$$S = f_1(N_d) \cdot f_2(N_w) \cdot f_3(N_c) \cdot (w_1 \cdot S^*_{BOW} + w_2 \cdot S^*_{OD}) \tag{4.6}$$

Đặt ROI_k là tập các vị trí điểm ảnh thuộc ảnh mẫu truy vấn thứ k. OPR_j là tập các điểm thuộc ảnh thứ j của một đoạn video trong database. Khi đó, N_d, N_w và N_c được tính dựa trên các công thức như sau:

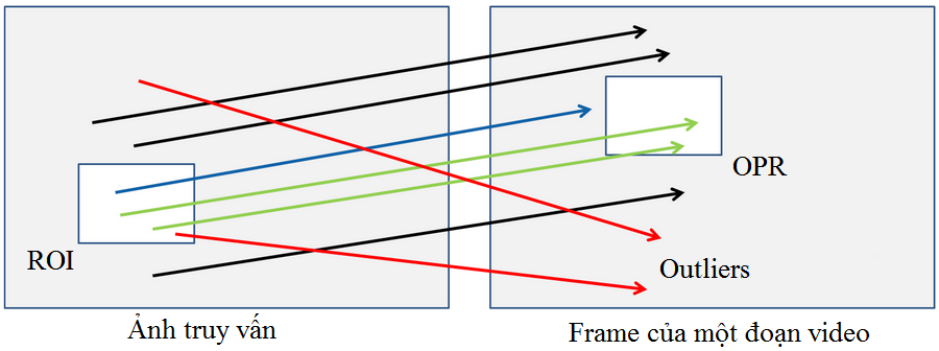
$$N_d = \sum_{k=1}^{n'} \sum_{j=1}^{m} card(\{(p,q)|p \in ROI_k, q \in OPR_j\}). \tag{4.7}$$

$$N_w = \sum_{k=1}^{n'} \sum_{j=1}^{m} card(\{(p,q)|p \notin ROI_k, q \in OPR_j\}). \tag{4.8}$$

$$N_c = \sum_{k=1}^{n'} \sum_{j=1}^{m} card(\{(p,q)|p \notin ROI_k, q \notin OPR_j\}). \tag{4.9}$$

trong đó, p và q là các điểm đặc trưng có cùng giá trị word ID đã được kiểm tra ràng buộc hình học. Chúng tôi giả sử rằng, các hàm số trên thuộc cùng một lớp hàm đa thức. Chúng tôi đề xuất ba hàm trọng số như sau:

$$\begin{cases} f_1(n) = 1 + n^2, \\ f_2(n) = 1 + n, \\ f_3(n) = 1 + \sqrt{n}. \end{cases}$$



Hình 4.2: Bốn loại cặp đặc trưng khai thác thông tin vị trí .

Công thức tính giá trị tương đồng cuối cùng trở thành:

$$S = (1 + N_d^2)(1 + N_w)(1 + \sqrt{N_c})(w_1 \cdot S^*_{BOW} + w_2 \cdot S^*_{OD}) \tag{4.10}$$

4.5.1 So sánh với các phương pháp state-of-the-art

Bảng 4.5.1 trình bày kết quả khi so sánh với các phương pháp state-of-the-art trên hai tập truy vấn. Phương pháp đề xuất cao hơn một cách đáng kể so với các phương pháp khác mặc dù chỉ sử dụng duy nhất một cặp bộ phát hiện và mô tả đặc trưng.

Bảng 4.4: So sánh các phương pháp trên tập dữ liệu INS2013 và INS2014.

Hướng tiếp cận	Phương pháp	INS2013	INS2014
Đặc trưng cục bộ	BOW	28.92	25.01
	DPM re-ranking	19.55	21.23
	BOW+DPM	32.18	28.21
	Multi-features[21]	31.33	28.77
	HE+WGC[38]	26.51	24.34
	PSR[97]	34.58	30.44
	TC[95]	20.50	N/A
Đặc trưng học sâu	FRCNN	21.60	20.67
	FRCNN+R+QE[8]	33.9	N/A
	BLCF[26]	32.3	N/A
Phương pháp đề xuất	**BOW+FRCNN+RANS**	**35.42**	**32.49**

4.6 Kết luận

Trong chương này chúng tôi trình bày hướng tiếp cận mới cho bài toán truy vấn đối tượng thị giác trong dữ liệu video lớn, bao gồm: (i) Đề xuất phương pháp reranking mới bằng cách kết hợp mô hình BOW với thông tin vị trí của đối tượng ứng viên sử dụng thuật toán phát hiện đối tượng; (ii)Đề xuất công thức tính score mới dựa trên thông tin về các visual word chung và thông tin vị trí của đối tượng ứng viên.

Chương 5
Mô hình kết hợp đặc trưng BOW với Deep Feature cho bài toán tìm kiếm người tại một địa điểm cho trước

5.1 Mở đầu

5.1.1 Nhu cầu thực tế

Ở hai chương trước, chúng tôi đã đề cập đến hai bài toán truy vấn trên cảnh vật và truy vấn một đối tượng duy nhất. Trong một số tình huống, người dùng có thể quan tâm nhiều hơn một đối tượng cùng lúc. Ví dụ như tìm người tại một địa điểm cho trước. Mặt khác, hai đối tượng mà người dùng quan tâm tìm kiếm nhiều nhất trên kho dữ liệu ảnh là người và địa điểm. Bài toán đặt ra là làm sao có thể tìm được ảnh có chứa người cần tìm tại một địa điểm cho trước. Do đó đây là loại thể thức truy vấn quan trọng và có tiềm năng ứng dụng to lớn. Có thể kể đến một số ứng dụng trong thực tế như: các hệ thống giám sát, quản lý dữ liệu video cá nhân, gợi nhớ quá khứ hỗ trợ trong việc điều trị chứng đãng trí. Tuy nhiên, đây cũng là một thể thức truy vấn khó bởi vì có rất nhiều biến thể khác nhau của các đối tượng được quan tâm như: kích thước, điều kiện ánh sáng, sự thay đổi về hình dáng của đối tượng theo trục thời gian.

5.1.2 Hướng tiếp cận của luận án

Trong chương này, chúng tôi đề xuất kết hợp với đặc trưng ngữ nghĩa để sàng lọc lại những đoạn video quay tại địa điểm không liên quan. Đối với người được truy vấn, trong trường hợp hướng mặt về phía camera, bài toán này tương đương với bài toán nhận diện gương mặt. Thay vì sử dụng đặc trưng VGG-Face vốn được thiết kế để sử dụng với độ đo khoảng cách chuẩn L_2, chúng tôi đề xuất sử dụng với máy phân lớp đặc trưng sử dụng nhân tuyến tính (linear kernel). Trong trường hợp người cần tìm không hướng mặt về phía camera, khi đó các đặc trưng nhận dạng dựa vào mặt không thể thực hiện được. Do trong cùng một cảnh quay, người đó không thể di chuyển quá nhanh để ra khỏi phạm vi của camera nên chúng tôi đề xuất theo vết dựa trên cảnh

(scene tracking) để làm tăng độ chính xác kết quả truy vấn. Hình 5.1 minh họa ý tưởng chính của phương pháp theo vết dựa trên cảnh quay này.

5.2 Tổng quan về hệ thống

Hệ thống được đề xuất của chúng tôi bao gồm 4 phần chính: truy vấn địa điểm dựa trên mô hình BOW, kiểm tra địa điểm dựa trên đặc trưng deep feature, nhận diện gương mặt và tổng hợp cuối cùng sử dụng theo vết địa điểm. Hình **??** phác thảo các luồng xử lý chính của hệ thống đề xuất. Toàn bộ quy trình thực hiện của thuật toán đề xuất được mô tả ở Thuật toán 5.

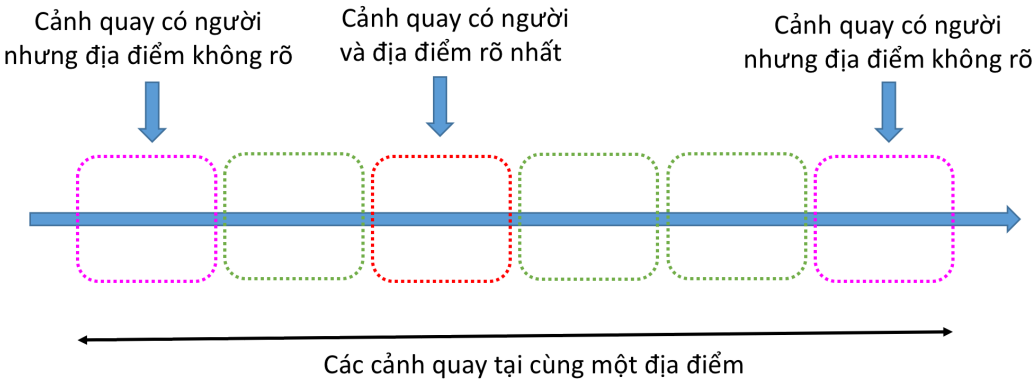
Thuật toán 5: Thuật toán tìm kiếm với truy vấn dạng hỗn hợp.

result = **CompoundQueryRetrieval** *($query_{loc}$, $query_{per}$, $database$, K)*

Đầu vào: $query_{loc}$: ảnh truy vấn mẫu của địa điểm của người cần quan tâm
$query_{per}$: ảnh truy vấn mẫu của người cần quan tâm
$database$: cơ sở dữ liệu ảnh cần truy vấn
K: số lượng ảnh/shot kết quả khi truy vấn theo địa điểm

Đầu ra : $result$: danh sách các ảnh/shot đã được sắp theo thứ tự giảm dần về độ tương đồng.

Bước 1: Xây dựng vector BOW trên truy vấn theo địa điểm
$loc_{BOW} \leftarrow$ BOW($query_{loc}$)
Bước 2: Rút trích đặc trưng gương mặt sử dụng đặc trưng học sâu
$pos\_face \leftarrow$ RemoveNoisyFace($query_{per}$)
$pos\_face_{VGG} \leftarrow$ VGG-Face(pos_face)
Bước 3: Huấn luyện mô hình phát hiện gương mặt với SVM
$neg\_face \leftarrow$ RetrieveSecondBestFace($face_{VGG}$)
$neg\_face_{VGG} \leftarrow$ VGG-Face(neg_face)
$face\_model \leftarrow$ TrainSVM(pos_face_{VGG}, neg_face_{VGG})
Bước 4: Tìm kiếm địa điểm dựa trên mô hình BOW
$loc\_shots \leftarrow$ RetrieveLocation($database$, K)
Bước 5: xếp hạng lại dựa trên đặc trưng học sâu gương mặt
$shot\_face_{VGG} \leftarrow$ VGG-Face(loc_shots)
$init\_resuls \leftarrow$ FaceScoringWithSVM($shot\_face_{VGG}$, $face\_model$)
Bước 6: Theo vết địa điểm
$result \leftarrow$ SceneTracking($init\_result$)
return $result$



Hình 5.1: Phương pháp xếp hạng lại dựa trên phương pháp theo vết địa điểm.

5.3 Thí nghiệm

5.3.1 Dữ liệu thí nghiệm

Trong chương này, chúng tôi sử dụng tập dữ liệu TRECVID INS với tập truy vấn INS2016 bao gồm các cặp đối tượng người và địa điểm được quan tâm. Để đánh giá hiệu quả của hệ thống chúng tôi sử dụng độ đo Mean Average Precision.

5.3.2 Độ chính xác và trực quan hóa kết quả truy vấn

Chúng tôi đánh độ chính xác của hệ thống đề xuất với các cấu hình sau:

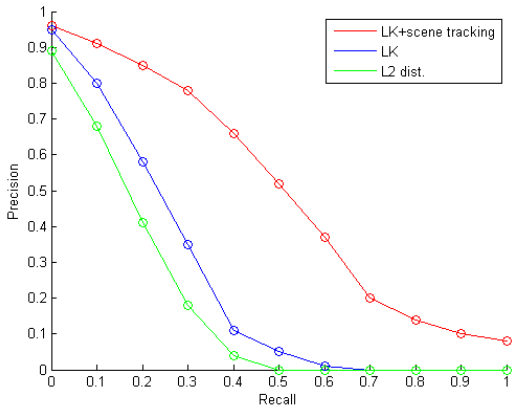
- Baseline: Sử dụng hệ thống framework của chúng tôi, sau bước kiểm tra địa điểm truy vấn sử dụng đặc trưng deep feature, chúng tôi tiến hành xếp hạng lại trên top K sử dụng độ đo khoảng cách L_2 làm độ đo chính để so sánh mặt.
- Linear Kernel: tương tự như hệ thống baseline nhưng chúng tôi sử dụng máy phân lớp với nhân tuyến tính để học mô hình gương mặt và so sánh với các gương mặt ứng viên.
- Linear Kernel+scene tracking: tương tự như Linear Kernel, nhưng chúng tôi còn áp dụng phương pháp theo vết địa điểm để xử lý những shot mà đối tượng cần tìm không quay mặt lại về phía camera.

Bảng 5.1 tóm tắt một số kết quả khi tiến hành trên các phương pháp khác nhau, sử dụng độ đo MAP. Kết quả cho thấy sử dụng máy phân lớp với nhân tuyến tính cho kết quả tốt hơn so với cấu hình baseline trong đó chỉ sử dụng độ đo L_2, cụ thể là đã

cải tiến độ chính xác từ 19.8% lên 25.9%. Hơn thế nữa, kết hợp với phương pháp theo vết địa điểm, độ chính xác của hệ thống tăng lên đáng kể từ 25.9% lên 50.6%.

Bảng 5.1: Kết quả thực nghiệm nghiệm trên tập dữ liệu TRECVID INS 2016.

Run	MAP
Linear Kernel + scene tracking	50.6
Linear Kernel	25.9
L_2 distance	19.8



Hình 5.2: Đường Precision-recall khi tiến hành thí nghiệm trên tập INS 2016.

Cần chú ý rằng, phương pháp theo viết địa điểm không chỉ giữ độ chính xác cao mà còn làm tăng độ phủ của thuật toán một cách đáng kể so với phương pháp sử dụng máy phân lớp. Điều này được thể hiện ở Hình 5.2, trong đó đường của phương pháp đề xuất *Linear Kernel+scene tracking* cao hơn đáng kể so với các đường còn lại.

5.4 Kết luận

Lấy cảm hứng từ sự thành công của các kỹ thuật học sâu (deep learning) trong những năm gần đây, chúng tôi cố gắng khai thác sức mạnh của đặc trưng deep feature cho bài toán tìm kiếm đối tượng. Cụ thể là chúng tôi đã đề xuất framework trong đó kết hợp điểm mạnh của mô hình BOW và đặc trưng deep feature phục vụ cho loại đối tượng mới của bài toán tìm kiếm đối tượng: tìm kiếm người tại địa điểm cho trước.

Chương 6
Tìm kiếm ảnh với mô tả truy vấn bằng ngữ nghĩa

6.1 Mở đầu

6.1.1 Nhu cầu thực tế

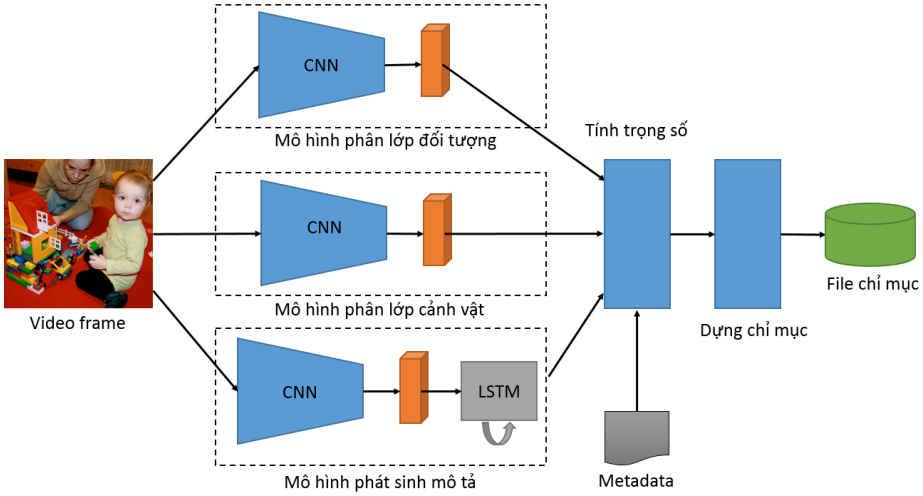
Trong chương này chúng tôi quan tâm giải quyết đến bài toán truy vấn với thể thức là câu mô tả: sử dụng từ ngữ mang tính chất mô tả cảnh vật, đối tượng cần tìm. Phương pháp truyền thống được các công cụ tìm kiếm sử dụng phổ biến hiện nay là dựa trên từ khoá mà người dùng mô tả đi kèm với ảnh, video. Các dữ liệu gán nhãn bởi người dùng này được gọi là metadata. Tuy nhiên, dữ liệu metadata không phải lúc nào cũng đầy đủ. Chương này tiếp cận bài toán theo hướng khai thác các khái niệm trong các cảnh quay video bằng cách sử dụng dữ liệu có sẵn từ trang web hoặc kho dữ liệu chung mà không cần gán nhãn bởi người dùng. Khái niệm *không gán nhãn* ở đây được hiểu là không thực hiện các thao tác chú thích, mô tả một cách thủ công và có *chủ đích phục vụ cho việc truy vấn.*

6.1.2 Hướng tiếp cận của luận án

Chương này tiếp cận theo hướng gán phụ đề ảnh dày đặc (dense captioning) từ tập dữ liệu gán nhãn có sẵn. Các mô hình gán phụ đề cũng sử dụng các mạng CNN nhưng **mở rộng với các phần tử LSTM (LSTM Cell)** giúp **mô tả quan hệ giữa các đối tượng chính trong ảnh**. Ngoài ra, do đối tượng mà người dùng quan tâm trong ảnh là không biết trước nên chúng tôi tiếp cận theo hướng **kết hợp các mô hình rút trích thuộc tính** từ các tập dữ liệu có số lượng nhãn lớn như MIT Places[12], Sun Attribute[27].

6.2 Mô hình tìm kiếm đối tượng với truy vấn biểu diễn bằng ngữ nghĩa

Hình 6.1 thể hiện chi tiết các bước xử lý rút trích đặc trưng các đối tượng chính trong ảnh, thuộc tính về mặt không gian, phát sinh câu mô tả, kết hợp với dữ liệu metadata để xây dựng file chỉ mục. Mỗi mạng CNN trong mô hình được huấn luyện từ các tập dữ liệu khác nhau và có thể thay đổi kiến trúc nền dễ dàng.



Hình 6.1: Chi tiết kiến trúc mạng rút trích các đặc trưng và đánh chỉ mục.

Rút trích đặc trưng ngữ nghĩa. Đây là bước xử quan trọng nhằm xác định các đặc trưng ngữ nghĩa chính có trong ảnh. Trái với đặc trưng cấp thấp, đặc trưng cấp cao có số chiều biểu diễn ít hơn, điều này giúp tiết kiệm chi phí lưu trữ và tính toán trong quá trình rút trích đặc trưng. Hơn thế nữa đặc trưng cấp cao có biểu diễn gần với ngôn ngữ biểu diễn truy vấn của người dùng hơn so với đặc trưng cấp thấp. Trong luận án này, đặc trưng nghĩa được chúng tôi đề xuất sử dụng bao gồm:

- Các đối tượng chính: chúng tôi đề xuất trích xuất 5 đối tượng có tín hiệu kích hoạt đầu ra cao nhất trong tập dữ liệu ImageNet 2014.
- Các thuộc tính không gian: bao gồm các thông tin phân loại không gian của ảnh/video frame lấy từ tập dữ liệu MIT Places và Sun Attribute.
- Quan hệ giữa các đối tượng trong ảnh: để biểu diễn tất cả các khía cạnh quan hệ của các đối tượng trong ảnh, chúng tôi sử dụng hướng tiếp cận mô tả phụ đề dày đặc (dense captioning) lấy từ tập dữ liệu Visual Genome[83].

- Dữ liệu metadata từ người dùng: đây là dữ liệu do người dùng tạo ra bao gồm: tiêu đề của ảnh/video, tóm tắt nội dung, nhãn (tag).

Xây dựng chỉ mục ngược. Sau khi rút trích các đặc trưng ngữ nghĩa, việc tìm kiếm bây giờ tương đương với việc so khớp trên văn bản. Do đó nhiệm vụ chính của bước này là tiến hành tạo chỉ mục cho các đặc trưng ngữ nghĩa rút trích từ các mạng Deep Neural Network.

Độ tương đồng giữa truy vấn và đoạn video. Trong phạm vi chương này, với dữ liệu thí nghiệm được xây dựng dựa trên đoạn (shot), do đó kết quả trả về là bảng xếp hạng của các đoạn video. Thuật toán 6 trình bày mã giả của thuật toán tìm kiếm video dựa trên mô tả bằng ngữ nghĩa.

Thuật toán 6: Thuật toán tìm kiếm với truy vấn dạng mô tả.

result = **AdhocQueryRetrieval**
  Đầu vào: D: tập hợp các shot video.
    Q: chuỗi câu mô tả truy vấn được cung cấp bởi người sử dụng.
  Đầu ra : *result*: File chỉ mục ngược lưu trữ thông tin đặc trưng ngữ nghĩa tổng hợp.
  Bước 1: Rút trích đặc trưng ngữ nghĩa đối tượng.
  Bước 2: Rút trích đặc trưng ngữ nghĩa mô tả khung cảnh.
  Bước 3: Rút trích đặc trưng ngữ nghĩa mô tả quan hệ giữa các đối tượng.
  Bước 4: Rút trích dữ liệu metadata.
  Bước 5: Tính trọng số dựa trên sự tương đồng ngữ nghĩa của các mô hình.
  Bước 6: Đánh chỉ mục.
  return *result*

6.3 Kết quả thử nghiệm

6.3.1 Dữ liệu thử nghiệm

Chúng tôi tiến hành thử nghiệm hệ thống đề xuất trên tập dữ liệu lớn TRECVID Ad-hoc Video Search (AVS). Tập dữ liệu bao gồm 4596 video được thu thập từ trên mạng internet với 144GB về kích thước và 600 giờ về thời lượng. Có tất cả 30 câu truy vấn với những nội dung mô tả không biết trước.

6.3.2 Kết quả thử nghiệm

Chúng tôi tiến hành so sánh với một số phương pháp truy vấn khác như:

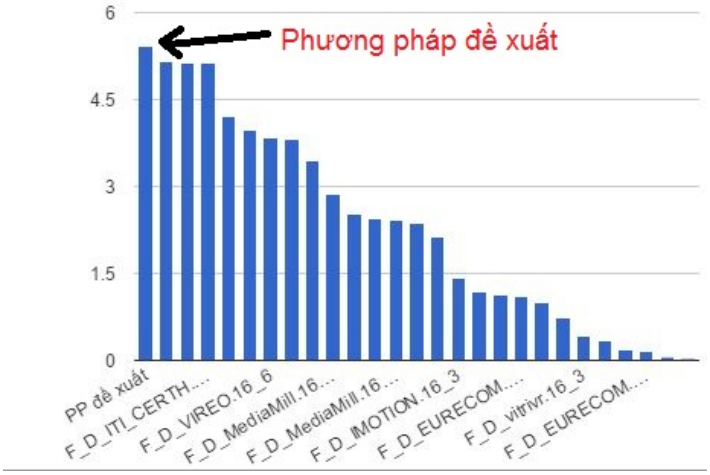
- ITI-CERTH: sử dụng bộ phát hiện khái niệm bằng thuật toán SVM kết hợp với phương pháp phân tích ngôn ngữ học (Linguistic Analysis) cho câu truy vấn với độ đo khoảng cách Histogram Intersection.
- HCR (Highlighted Concept Reranking) [33]: sử dụng khoảng 10.000 khái niệm do chính các tác giả xây dựng, kết hợp với các dataset ImageNet, MIT Places và sử dụng tri thức của chuyên gia để lọc bớt khái niệm không quan trọng.

Bảng 6.1 so sánh độ chính xác của phương pháp đề xuất so với hai phương pháp ITI-CERTH và HCR trong hạng mục truy vấn bằng câu mô tả tự động. Kết quả quả cho thấy phương pháp đánh trọng số các khái niệm bằng phương pháp *tf-idf* kết hợp với dữ liệu metadata cho độ chính xác cao hơn so với hai phương pháp còn lại.

Bảng 6.1: Kết quả thực nghiệm nghiệm trên tập dữ liệu TRECVID INS 2016.

Phương pháp	MAP
ITI-CERTH	5.1
HCR [33]	4.6
Phương pháp đề xuất	**5.4**

Hình 6.2 thể hiện độ chính xác của tất cả các nhóm nghiên cứu tham gia trong cùng một hạng mục xử lý tự động của AVS. Trục ngang là mã viết tắt của tất cả các nhóm, trục đứng là độ chính xác tính theo độ đo MAP.



Hình 6.2: So sánh độ chính xác với các phương pháp khác tại TRECVID AVS 2016.

Từ các hướng tiếp cận của các nhóm nghiên cứu ta có thể thấy rằng, việc sử dụng kết hợp các đặc trưng ngữ nghĩa đến từ việc khai thác các khái niệm của các dataset

còn rất nhiều tiềm năng. Độ chính xác của các hệ thống trên hạng mục còn thấp cho thấy đây là bài toán khó và còn rất nhiều khả năng mở rộng phát triển. Mặt khác, mặc dù dữ liệu metadata là không đầy đủ và đôi khi có chứa nhiễu nhưng việc khai thác nguồn dữ liệu này một cách hợp lý cũng góp phần làm tăng độ chính xác của hệ thống. Điều này thể hiện thông qua việc hệ thống đề xuất có độ chính xác cao hơn hai hệ thống không khai thác nguồn dữ liệu này là ITI-CERTH và HCR.

6.4 Kết luận

Chương này đề cập đến loại truy vấn mới sử dụng mô tả bằng ngữ nghĩa với các khía cạnh khác nhau của một khung ảnh. Chúng tôi sử dụng các nhãn liên quan đến đối tượng chính, các thuộc tính liên quan đến không gian, quan hệ giữa các đối tượng trong ảnh và cả dữ liệu metadata cung cấp bởi người dùng và kết hợp lại với nhau để xây dựng hệ thống truy vấn. So sánh với các phương pháp được đề xuất trong cuộc thi TRECVID AVS 2016 cho thấy, ở hạng mục tìm kiếm tự động phương pháp đề xuất của chúng tôi cho độ chính xác cao nhất.

Chương 7
Kết luận

7.1 Những kết quả đã đạt được

Trong luận án này chúng tôi đề xuất một số phương pháp và hệ thông để cải tiến bài toán tìm kiếm đối tượng. Trong Chương 3, chúng tôi đề xuất thuật toán kết hợp cấu trúc chỉ mục ngược với spatial pyramid (kim tự tháp không gian) để tăng tốc độ và độ chính xác của hệ thống tìm kiếm trên đối tượng lớn có nhiều đặc trưng hoa văn. Trong Chương 4, chúng tôi đề xuất hai thuật toán để cải tiến độ chính xác trong quá trình tìm kiếm với đối tượng nhỏ, ít đặc trưng hoa văn. Phương pháp đầu tiên là sử dụng mạng neural network để xác định giá trị trọng số kết hợp hai mô hình BOW và thuật toán phát hiện đối tượng. Phương pháp tiếp theo, chúng tôi khai thác vị trí tương đối giữa các visual word với vị trí đối tượng đề xuất bởi thuật toán phát hiện đối tượng. Mỗi loại visual word sẽ có một mức độ tin cậy riêng được biển diễn bởi các hàm trọng số tương ứng. Trong Chương 5, chúng tôi giải quyết bài toán tìm kiếm với loại truy vấn mới: tìm người tại một địa điểm cho trước. Giải pháp được đề xuất bao gồm kết hợp các đặc trưng của mô hình BOW với đặc trưng rút trích từ lớp cuối cùng của mạng CNN huấn luyện sẵn. Ngoài ra để tăng độ chính xác, chúng tôi huấn luyện các đặc trưng với máy phân lớp sử dụng linear kernel. Ngoài ra, để tăng độ phủ của hệ thống, chúng tôi đề xuất phương pháp theo vết trên địa điểm. Chương 6 chúng tôi giải quyết bài toán tìm kiếm dựa trên mô tả ngữ nghĩa của người dùng (dạng văn bản). Để giải quyết bài toán này, chúng tôi đề xuất mô tả mỗi frame video bằng tất cả những khái niệm trong các tập dataset hiện có. Câu truy vấn trước khi người dùng đưa vào sẽ được chuẩn hóa nhằm giảm bớt sự sai lệch do yếu tố ngôn ngữ gây ra.

Trong quá trình thực hiện luận án này, ngoài các bài báo đã được công bố trong các hội nghị, tạp chí quốc tế có uy tín, tác giả còn đạt được những giải thưởng liên quan đến các công trình đã nghiên cứu như:

- Hạng nhất cuộc thi TRECVID Instance Search (INS) 2014 cho hạng mục hệ thống truy vấn tự động, hạng nhì cho các năm 2015 và 2016.
- Top 10% bài báo xuất sắc của hội nghị MMSP 2015

- Hạng nhất cuộc thi SHREC 2016 cho hạng mục tìm kiếm đối tượng 3D với truy vấn dạng bán phần (SHREC 2016 Track on Partial Shape Queries for 3D Object Retrieval)
- Hạng nhất cuộc thi TRECVID Ad-hoc Video Search (AVS) 2016 cho hạng mục hệ thống truy vấn tự động

7.2 Một số hướng phát triển luận án

Dưới đây là một số hướng phát triển cho một số thể thức và loại đối tượng truy vấn:

Đối tượng nhỏ ít đặc trưng: chúng tôi đề xuất phát triển trong tương lai là khai thác cấu trúc chỉ mục ngược cho bài toán phát hiện đối tượng. Với hướng tiếp cận này, việc phát hiện đối tượng có thể được thực hiện trên dữ liệu lớn với thời gian gần với thời gian thực.

Nhóm đối tượng: hướng tiếp cận mà chúng tôi đề nghị phát triển bao gồm: tích hợp cấu trúc chỉ mục ngược và chiến lược nhánh cận trong việc lưu trữ đặc trưng gương mặt người rút trích từ kho dữ liệu ảnh. Khi tiến hành so sánh các vector đặc trưng trên cấu trúc chỉ mục ngược, nếu khoảng cách vượt quá một ngưỡng cho trước thì sẽ không xử lý tiếp trên những thành phần còn lại của vector đặc trưng. Ngoài ra, mạng CNN dùng để rút trích đặc trưng gương mặt sẽ được huấn luyện lại để thích nghi với gương mặt của đối tượng truy vấn mới.

Truy vấn dựa trên ngữ nghĩa: chúng tôi đề xuất hướng tiếp cận kết hợp với các công cụ tìm kiếm hình ảnh dựa trên văn bản hiện nay để học online và cho kết quả gần với dữ liệu gán nhãn trước đó. Hướng tiếp cận này có thể hiểu là dựa trên những biểu diễn của những từ đồng nghĩa hoặc gần nghĩa.

Phụ lục A

Các công trình đã công bố

Tạp chí quốc tế:

[CT1] Vinh-Tiep Nguyen, Thanh Duc Ngo, Minh-Triet Tran, Duy-Dinh Le, Duc Anh Duong: A Combination of Spatial Pyramid and Inverted Index for Large-Scale Image Retrieval, tạp chí International Journal of Multimedia Data Engineering and Management, Volume 6, Number 2, trang 37-51, năm 2015, ISSN: 1947-8534.

[CT2] Vinh-Tiep Nguyen, Thanh Duc Ngo, Minh-Triet Tran, Duy-Dinh Le, Duc Anh Duong: Persons-In- Places: a Deep Features Based Approach for Searching a Specific Person in a Specific Location, Informatica2017, Volume 41, Number 2, trang 149–158, năm 2017.

[CT3] Vinh-Tiep Nguyen, Duy Dinh Le, Minh-Triet Tran, Tam V. Nguyen, Thanh Duc Ngo, Shinichi Satoh, Duc Anh Duong: Video Instance Search via Spatial Fusion of Visual Words and Object Proposals, International Journal of Multimedia Information Retrieval, 2019 (được chấp nhận đăng ngày 15 tháng 4 năm 2019).

Hội nghị quốc tế:

[CT4] Vinh-Tiep Nguyen, Thanh Duc Ngo, Duy-Dinh Le, Minh-Triet Tran, Duc Anh Duong, Shinichi Satoh: Semantic Extraction and Object Proposal for Video Search, International Conference on Multimedia Modeling (MMM), 2017, Reykjavik, Iceland.

[CT5] Vinh-Tiep Nguyen, Minh-Triet Tran, Thanh Duc Ngo, Duy Dinh Le, Duc Anh Duong: Searching a specific person in a specific location using deep features, the Seventh Symposium on Information and Communication Technology (SoICT), 2016, Ho Chi Minh city, Vietnam.

[CT6] Vinh-Tiep Nguyen, Khanh-Duy Le, Minh-Triet Tran, Morten Fjeld: NowAndThen: a Social Network-Based Photo Recommendation Tool Supporting Reminiscence, International Conference on Mobile and Ubiquitous Multimedia (MUM), 2016, Rovaniemi, Finland.

[CT7] Vinh-Tiep Nguyen, Dinh-Luan Nguyen, Minh-Triet Tran, Duy-Dinh Le, Duc Anh Duong, Shinichi Satoh: Query-adaptive late fusion with neural network for instance search, MMSP 2015: 1-6 (Top 10% Paper Award)