BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Nguyễn Thị Lan Phương

MỘT SỐ KỸ THUẬT NÂNG CAO HIỆU QUẢ TRA CỨU ẢNH THEO NỘI DUNG DỰA TRÊN ĐỘ ĐO KHOẢNG CÁCH THÍCH NGHI VÀ PHÂN CỤM PHỔ

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Hà Nội - 2023

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Nguyễn Thị Lan Phương

MỘT SỐ KỸ THUẬT NÂNG CAO HIỆU QUẢ TRA CỨU ẢNH THEO NỘI DUNG DỰA TRÊN ĐỘ ĐO KHOẢNG CÁCH THÍCH NGHI VÀ PHÂN CỤM PHỔ

LUẬN ÁN TIẾN SĨ KHOA HỌC MÁY TÍNH

Mã số: 9 48 01 01

Xác nhận của Học viện Khoa học và Công nghệ

Người hướng dẫn 1

PGS.TS. Ngô Quốc Tạo

Người hướng dẫn 2

TS. Nguyễn Ngọc Cương

Hà Nội - 2023

LỜI CAM ĐOAN

Nghiên cứu sinh xin cam đoan đây là công trình nghiên cứu của riêng nghiên cứu sinh. Các kết quả được viết chung với các tác giả khác đều được sự đồng ý của các đồng tác giả trước khi đưa vào luận án. Các kết quả được trình bày trong luận án là mới, các số liệu là trung thực và chưa từng được ai công bố trong các công trình nào khác./.

Nghiên cứu sinh

Nguyễn Thị Lan Phương

LỜI CẢM ƠN

Luận án này được thực hiện tại Học viện Khoa học và Công nghệ, Viện Hàn Lâm khoa học và Công nghệ Việt Nam dưới sự hướng dẫn, chỉ bảo tận tình của PGS.TS. Ngô Quốc Tạo và TS. Nguyễn Ngọc Cương, những người mà từ đó Nghiên cứu sinh đã học được rất nhiều điều quý báu, các thầy là tấm gương sáng cho nghiên cứu sinh trong nghiên cứu chuyên môn cũng như trong cuộc sống. Nghiên cứu sinh xin gửi lời cảm ơn sâu sắc đến các thầy cô tại Học viện Khoa học và Công nghệ, Viện Hàn Lâm khoa học và Công nghệ Việt Nam về sự giúp đỡ, chỉ dẫn tận tình trong quá trình nghiên cứu và hoàn thành luận án.

Nghiên cứu sinh xin gửi lời cảm ơn đến PGS. TS. Nguyễn Đức Dũng, PGS. TS. Nguyễn Hữu Quỳnh, TS. Đào Thúy Quỳnh đã có nhiều góp ý chuyên môn và sự động viên tinh thần giúp vượt qua nhiều khó khăn trong quá trình nghiên cứu cũng như trong cuộc sống. Nghiên cứu sinh cũng xin gửi lời cảm ơn các thầy giáo, cô giáo ở Viện Công nghệ thông tin, Viện Hàn lâm khoa học và Công nghệ Việt Nam đã tạo điều kiện thuận lợi và giúp đỡ trong thời gian học tập tại Viện. Nghiên cứu sinh xin chân thành cảm ơn Hội đồng đánh giá luận án tiến sĩ cấp cơ sở đã góp ý những ý kiến quý báu để Nghiên cứu sinh sửa bản luận án được tốt nhất. Nghiên cứu sinh cũng xin gửi lời cảm ơn tới lãnh đạo Phân hiệu Đại học Thái Nguyên tại tỉnh Lào Cai và các đồng nghiệp.

Cuối cùng tác giả xin bày tỏ lòng biết ơn đến gia đình và bạn bè đã động viên giúp đỡ về tinh thần, thời gian để hoàn thành luận án.

Hà Nội, ngày tháng 10 năm 2023

Nghiên cứu sinh

Nguyễn Thị Lan Phương

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

TBIR	Text-based image retrieval	Tra cứu ảnh dựa trên văn bản
CBIR	Content-based image retrieval	Tra cứu ảnh dựa trên nội dung
IRIC	Image retrieval method using Incremental clustering	Phương pháp tra cứu ảnh sử dụng phân cụm tăng dần
CISE	Clustering Images Set using Eigenvectors	Nhóm ảnh được thiết lập bằng cách sử dụng Eigenvectos
INC	Incremental Clustering	Phân cụm tăng dần
CNN	convolutional neural networks	Mạng nơ ron phức hợp
ODLDA	Image Retrieval using the optimal distance and linear Discriminant analysis	Tra cứu ảnh bằng cách sử dụng khoảng cách khoảng cách tối ưu và phân tích phân biệt tuyến tính
OASIS	Online Algorithm for Scalable Image Similarity	Mở rộng thuật toán trực tuyến cho sự giống nhau của ảnh
DML	Distance metric learning	Học khoảng cách khoảng cách
DCA	Discriminative Components Analysis	Phân tích các thành phần phân biệt
IR	Information retrieval	Tra cứu thông tin
RF	Relevance feedback	Mức độ trả lời liên quan
ST	Semantic template	Mẫu ngữ nghĩa
RGB	Red Green Blue	Đỏ lục lam
CCVs	Color coherence vectors	Các vectơ liên kết màu
SPCA	Shift-invariant principal component analysis	Phân tích thành phần chính thay đổi – bất biến

MLE	Maximum likelihood estimation	Tính toán khả năng xảy ra tối đa
SAR	Simultaneous autoregressive (SAR) model	Mô hình tự động hồi phục đồng thời
MRF	Markov random field	Trường ngẫu nhiên Markov
LSE	Least square error	Lỗi bình phương ít nhất
RISER	Rotation-invariant SAR	Bất biến xoay SAR
PWT	Pyramid-structured wavelet transform	Biến đổi Wavelet có cấu trúc kim tự tháp
TWT	Tree-structured wavelet transform	Biến đổi Wavelet có cấu trúc cây
KL	Kullback–Leibler	Kullback–Leibler
PCA	Principal component analysis	Phân tích thành phần chính
DCT	Discrete cosine transforms	Biến đổi cosin rời rạc
CWT	Complex wavelet transforms	Biến đổi Wavelet phức tạp

DANH MỤC CÁC BẢNG

DANH MỤC HÌNH VẼ

PHẦN MỞ ĐẦU

1. Tính cấp thiết của luận án

Trong thập kỷ qua, chúng ta đã chứng kiến sự tăng trưởng liên tục của số lượng ảnh kỹ thuật số được chụp, lưu trữ và chia sẻ mỗi ngày. Ước tính số lượng ảnh kỹ thuật số được chụp năm 2021 là hơn 5 nghìn tỷ. Khoảng 85% trong số đó là chụp bằng điện thoại di động. Một phần lớn trong số chúng có sẵn trên Internet thông qua các trang web , thư viện ảnh (Flickr và Shutterstock), và các phương tiện truyền thông xã hội khác nhau Facebook, Instagram.... Phần lớn các cơ sở dữ liệu ảnh này , không được sắp xếp cũng không đính kèm siêu dữ liệu và thẻ. Ngoài ra, cơ sở dữ liệu ảnh phổ biến trong các lĩnh vực ứng dụng như phòng chống tội phạm, y học, kiến trúc, viễn thám , …Các kỹ thuật thu truyền và lưu trữ ảnh ngày càng phát triển đã cho phép xây dựng các cơ sở dữ liệu ảnh khổng lồ. Tra cứu ảnh dựa vào nội dung (CBIR) giải quyết bài toán quản lý thư viện ảnh, phân loại ảnh, nhận dạng đối tượng trong ảnh, tra cứu hình ảnh trên mạng và nhiều ứng dụng khác liên quan đến xử lý ảnh và thị giác máy tính. Do vậy, việc tra cứu nhanh chóng và chính xác một bức ảnh mong muốn trong một cơ sở dữ liệu ảnh số lớn và đa dạng là một nhiệm vụ hết sức khó khăn, đầy thách thức trong lĩnh vực thị giác máy tính hiện nay.

Tra cứu ảnh dựa vào nội dung cho phép người dùng tra cứu hình ảnh dựa trên các đặc trưng cụ thể của ảnh như màu sắc, cấu trúc, hình dạng, điều này giúp cải thiện hiệu quả của việc tra cứu ảnh, đặc biệt khi người dùng không thể sử dụng từ khoá hoặc chưa biết chính xác mô tả của ảnh. Ngoài ra tra cứu ảnh dựa trên nội dung khắc phục ngôn ngữ hạn chế và ảnh chưa có gán nhãn đầy đủ có nghĩa là người dùng tra cứu ảnh bằng cách sử dụng ảnh mà người ta có sẵn thay vì phải biết chính xác từ khoá mô tả. Tra cứu ảnh dựa vào nội dung cho phép người dùng tra cứu các hình ảnh tương tự về nội dung hoặc phong cách, mô tả khả năng khám phá và khám phá theo các tiêu chí tương tự. Nghiên

cứu trong lĩnh vực tra cứu ảnh dựa vào nội dung thúc đẩy sự phát triển của các thuật toán xử lý ảnh, trí tuệ nhân tạo, học máy. Điều này có thể dẫn đến việc cải thiện hiệu suất và tính năng của các hệ thống tra cứu ảnh trong tương lai. Chính vì vậy rất nhiều nhà khoa học tập trung nghiên cứu nâng cao hiệu quả tra cứu ảnh dựa vào nội dung.

Tra cứu ảnh dựa trên nội dung (CBIR), còn được gọi là truy vấn theo nội dung ảnh (Query By Image Content -QBIC), là một kỹ thuật tự động lấy ảnh làm truy vấn và trả về một tập hợp ảnh tương tự [1, 2]. Kỹ thuật tra cứu ảnh dựa trên nội dung sử dụng trực quan nội dung của các ảnh được mô tả dưới dạng cấp thấp như màu sắc, kết cấu, hình dạng và vị trí không gian để thể hiện và tra cứu ảnh trong cơ sở dữ liệu. Hệ thống lấy các ảnh tương tự khi một ảnh mẫu hoặc phác họa làm đầu vào cho hệ thống. Ảnh truy vấn được chuyển đổi thành véc tơ biểu diễn đặc trưng ảnh bằng cách sử dụng cùng một phương pháp trích xuất đặc trưng được sử dụng để xây dựng cơ sở dữ liệu. Điều này giảm đáng kể những khó khăn của cách tiếp cận thuần túy dựa trên chú thích, bởi vì quá trình trích rút đặc trưng có thể thực hiện tự động, tốn ít thời gian.

Các thí nghiệm mở rộng trên các hệ thống tra cứu ảnh dựa trên nội dung cho thấy đặc trưng mức thấp thường không thể mô tả các khái niệm ngữ nghĩa mức cao trong tâm trí người dùng. Do đó, hiệu suất của tra cứu ảnh dựa vào nội dung vẫn còn xa so với mong đợi của người dùng. [2, 3, 4, 5, 6 , 7 , 8 , 9 , 10, 11, 12, 13] đề cập đến ba cấp độ truy vấn trong tra cứu ảnh dựa vào nội dung.

Cấp độ 1: Tra cứu nguyên thủy như màu sắc, kết cấu, hình dạng hoặc không gian vị trí của các yếu tố ảnh hưởng. Truy vấn thông thường là một truy vấn bằng ảnh mẫu.

Cấp độ **2:** Tra cứu các đối tượng được xác định bởi đăc trưng dẫn xuất với một số mức độ suy luận logic về tên của đối tượng được mô tả trong ảnh. Ví dụ, hãy tìm một bức tranh trong cảnh Sapa .

Cấp độ 3: Tra cứu theo các thuộc tính trìu tượng, liên quan đến một lượng lớn suy luận cấp cao về mục đích của đối tượng hoặc cảnh được mô tả. Điều này bao gồm tra cứu các sự kiện được đặt tên, các ảnh có ý nghĩa về cảm xúc hoặc tôn giáo , v.v. Ví dụ, để tìm ảnh của một đám đông vui vẻ thì tìm ảnh có mô tả độ vui vẻ. Để tìm mô tả độ vui vẻ thì căn cứ vào biểu cảm của người trong ảnh như cười mồm chữ O và mắt chữ A chẳng hạn.

Các phương pháp tra cứu anh cấp độ 2 và 3 cùng được gọi là tra cứu ảnh ngữ nghĩa, khoảng cách giữa cấp 1 và 2 là khoảng cách ngữ nghĩa. Cụ thể hơn, sự khác biệt giữa mô tả hạn chế của đặc trưng ảnh mức thấp và sự phong phú của ngữ nghĩa người dùng được gọi là khoảng cách ngữ nghĩa.

Người dùng trong tra cứu cấp 1 thường được yêu cầu gửi ảnh mẫu hoặc phác họa làm truy vấn. Tra cứu ảnh ngữ nghĩa thuận tiện hơn cho người dùng vì nó hỗ trợ truy vấn theo từ khóa hoặc theo kết cấu. Do đó, để nâng cao chất lượng tra cứu ảnh theo các khái niệm mức cao, một hệ thống tra cứu ảnh dựa vào nội dung cần cung cấp hỗ trợ đầy đủ trong việc thu hẹp khoảng cách ngữ nghĩa giữa ảnh số và sự phong phú của ngữ nghĩa của con người [13].

Các phương pháp tra cứu ảnh ở cấp độ 3 là khó khăn và vì ít phổ biến hơn và được ứng dụng trong các lĩnh vực cụ thể như bảo tàng nghệ thuật. thư viện v.v. Hiện nay, hệ thống tra cứu ảnh chủ yếu thực hiện ở cấp độ 2.

Trong cuộc khảo sát gần nhất Xing và cộng sự [14] đã phân loại các kỹ thuật tiên tiến trong giảm khoảng cách ngữ nghĩa thành năm loại: một là sử dụng bản thể đối tượng để xác định khái niệm mức cao; Hai là sử dụng các công cụ học máy để liên kết các đặc trưng mức thấp với các khái niệm truy vấn mức cao; Ba là đưa phản hồi liên quan (relevance feedback – RF) vào vòng lặp tra cứu cho học liên tục về ý định của người dùng [15]; Bốn là tạo mẫu ngữ

nghĩa (semantic template – ST) để hỗ trợ tra cứu ảnh mức cao; Năm là sử dụng cả nội dung trực quan của ảnh và thông tin văn bản thu được từ ảnh trên trang web.

Ngoài ra còn có Phương pháp học số liệu để nhận dạng[16].

Trong một hệ thống tra cứu ảnh dựa vào nội dung điển hình, các đặc trưng ảnh mức thấp bao gồm màu sắc, kết cấu và hình dạng, được tự động trích xuất và biểu diễn dưới dạng véc tơ đặc trưng. Cũng lưu ý ở đây là, các véc tơ đặc trưng được xem là tốt nếu chúng mang ý nghĩa ngữ nghĩa của ảnh và phục vụ tốt cho việc so sánh ảnh. Để tìm một số ảnh mong muốn, người dùng đưa vào hệ thống một ảnh mẫu và hệ thống trả về một danh sách các ảnh tương tự dựa trên nội dung của ảnh. Khi hệ thống đưa ra danh sách các ảnh tương tự với ảnh truy vấn, người dùng đánh dấu những ảnh phù hợp nhất với ảnh truy vấn để thu được một danh sách phản hồi. Hệ thống dựa vào danh sách phản hồi này để huấn luyện ra một mô hình để cải tiến độ chính xác tra cứu ảnh.

Do đó, biểu diễn của ảnh bằng véc tơ đặc trưng [17, 18] và so sánh độ tương đồng là hai yếu tố chính ảnh hưởng đến hiệu quả của tra cứu ảnh dựa trên nội dung. Nâng cao hiệu quả của hệ thống tra cứu ảnh dựa trên nội dung là một vấn đề thách thức trong nghiên cứu. Để nâng cao hiệu quả, NCS cần giảm khoảng cách ngữ nghĩa trong CBIR. Khoảng cách ngữ nghĩa hàm ý sự khác biệt giữa đặc trưng mức thấp và khái niệm ngữ nghĩa mức cao của ảnh. Để giảm khoảng cách ngữ nghĩa này, cần đưa một số mô hình học máy vào quá trình tra cứu hình ảnh.

Gần đây, có những kết quả tốt do việc áp dụng CNN cho CBIR. Nó đã chỉ ra rằng nếu CNN được huấn luyện theo cách giám sát đầy đủ trên một tập lớn các ảnh có nhãn CNN có thể giải quyết nhiều loại tác vụ như phân loại ảnh đối tượng, nhận dạng ảnh, phát hiện thuộc tính và tra cứu ảnh [4 , 19, 20, 21, 22, 23]. Nghiên cứu trong [22] đã chỉ ra rằng hiệu suất của các hệ thống CBIR sử dụng CNN có tính cạnh tranh ngay cả khi CNN được huấn luyện cho một nhiệm vụ phân loại. Để nâng cao hiệu quả ngay từ quá trình xây dựng bộ đặc trưng,

phương pháp được đề xuất sẽ tận dụng ưu điểm của đặc trưng được trích rút bởi CNN. Bên cạnh đó, phương pháp đề xuất sẽ kết hợp các kỹ thuật học tương tự để có một độ đo tương tự cải tiến.

2. Mục tiêu của luận án

Mục tiêu chung của luận án:

Đề xuất được phương pháp tra cứu ảnh cho nâng cao độ chính xác tra cứu.

Mục tiêu cụ thể của luận án:

- Cải tiến phương pháp tra cứu ảnh bằng phương pháp ODLDA thông qua tìm một phép đo khoảng cách tối ưu, mà giảm khoảng cách giữa các cặp ảnh có độ tương tự cao và tối đa hóa khoảng cách giữa các cặp ảnh có độ tương tự thấp.

- Đề xuất phương pháp tra cứu ảnh dựa trên lý thuyết cắt đồ thị, mà không phải tính ma trận Laplacian, các giá trị riêng và các véc tơ riêng.

3. Đối tượng nghiên cứu

Đối tượng nghiên cứu của luận án là tra cứu ảnh dựa trên nội dung bằng cách kết hợp khoảng cách tối ưu và phân tích phân biệt tuyến tính, tiến hành thực nghiệm trên tập cơ sở dữ liệu tập ảnh Corel (1 0.800 ảnh), phân hoạch đồ thị với cơ sở dữ liệu ảnh SIMPLIcity (1.000 ảnh với 10 chủ đề. Mỗi ảnh có kích thước 256×384 hoặc 384×256).

4. Phương pháp nghiên cứu của luận án

Phương pháp nghiên cứu của luận án là nghiên cứu lý thuyết và nghiên cứu thực nghiệm. Về nghiên cứu lý thuyết: giới thiệu về tra cứu ảnh dựa vào nội dung, một số nghiên cứu ảnh dựa vào nội dung, trích rút đặc trưng, thông tin không gian, đo khoảng cách, phân cụm, giảm khoảng cách ngữ nghĩa, phân tích phân biệt tuyến tính, đánh giá hiệu năng.

5. Bố cục của luận án

Luận án này được bố cục thành ba chương:

Chương 1: Tổng quan về tra cứu ảnh dựa trên nội dung.

Chương 2: Nâng cao hiệu quả của việc tra cứu ảnh dựa trên nội dung bằng cách kết hợp tối ưu khoảng cách và phân tích phân biệt tuyến tính.

Chương 3: Cải thiện hiệu quả của tra cứu ảnh dựa trên nội dung sử dụng phân hoạch đồ thị

Cuối cùng, luận án đưa ra một số kết luận và định hướng nghiên cứu trong tương lai.

6. Kết quả và tính mới của luận án

Đóng góp vào hướng nghiên cứu, luận án đưa ra được những đóng góp sau:

(1) Luận án nâng cao độ chính xác tra cứu ảnh thông qua việc xây dựng cơ sở dữ liệu véc tơ đặc trưng với mạng học sâu CNN AlextNet.

(2) Trong quá trình học độ đo tương tự, luận án xem xét cả tập liên quan và tập không liên quan và sử dụng phương pháp học phân tích phân biệt tuyến tính LDA để tiến hành điều chỉnh hàm trọng số của hàm khoảng cách.

(3) Đề xuất phương pháp tra cứu ảnh hiệu quả sử dụng phân hoạch đồ thị (An efficient image retrieval method using a graph clustering-MGC) mà khai thác đầy đủ thông tin độ tương tự của tập ảnh. Kết quả thực nghiệm của luận án trên cơ sở dữ liệu đặc trưng gồm 1.000 ảnh đã chỉ ra rằng phương pháp được đề xuất ***MGC*** cung cấp một độ chính xác cao hơn so với các phương pháp khác.

Chương 1. TỔNG QUAN VỀ TRA CỨU ẢNH DỰA VÀO NỘI DUNG

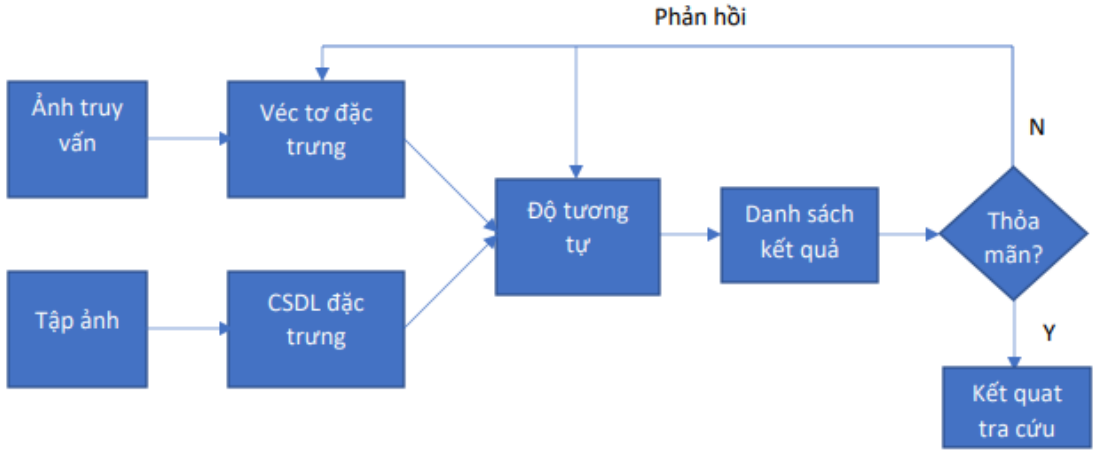
Chương này giới thiệu cơ bản về tra cứu ảnh dựa vào nội dung bao gồm: xem xét sự phát triển của kỹ thuật tra cứu ảnh dựa vào nội dung [24], cách mô tả nội dung trực quan, độ đo khoảng cách giữa các nội dung trực quan, lược đồ chỉ mục, tạo truy vấn, cơ chế phản hồi liên quan. Bên cạnh đó, chương này cũng trình bày về đánh giá hiệu năng hệ thống. Cuối cùng, chương này đưa ra một số kết luận và định hướng cho nghiên cứu.

1.1. Giới thiệu

Tra cứu ảnh dựa vào nội dung (CBIR) là sử dụng nội dung trực quan của ảnh để tìm những ảnh trong những cơ sở dữ liệu ảnh lớn mà tương tự với ảnh truy vấn. CBIR là một lĩnh vực nghiên cứu tích cực và phát triển nhanh chóng từ những năm 1990. Trong những thập kỹ qua, CBIR có có những tiến bộ về cả lý thuyết và ứng dụng, tuy nhiên, độ chính xác và tốc độ của các hệ thống CBIR vẫn cần tiếp tục được nghiên cứu cải tiến. Trước khi giới thiệu lý thuyết cơ bản về tra cứu ảnh dựa trên nội dung NCS giới thiệu sơ lược về sự phát triển của nó. Đầu tiên, nghiên cứu về tra cứu ảnh bắt đầu từ những năm 1970 đến 1979, hội nghị về kỹ thuật ứng dụng báo ảnh được tổ chức tại Florence. Kể từ đó, tiềm năng ứng dụng của kỹ thuật quản lý cơ sở dữ liệu ảnh đã thu hút sự quan tâm của các nhà nghiên cứu. Ban đầu, tra cứu ảnh sử dụng cách tiếp cận chú thích ảnh. Nói cách khác, ảnh đầu tiên được chú thích bằng văn bản và sau đó được tra cứu bằng cách tiếp cận dựa trên văn bản bởi hệ thống quản lý cơ sở dữ liệu truyền thống.

Tra cứu ảnh dựa trên nội dung (CBIR), sử dụng nội dung trực quan của ảnh như màu sắc, hình dạng, kết cấu và bố cục không gian để biểu diễn và lập chỉ mục cho hình ảnh. Trong các hệ thống CBIR điển hình (Hình I.1), nội dung trực quan của ảnh trong cơ sở dữ liệu được trích xuất và mô tả bằng các véc tơ đặc trưng đa chiều. Các véc tơ đặc trưng của các ảnh trong cơ sở dữ liệu được

trích rút để tạo ra một cơ sở dữ liệu đặc trưng. Để tra cứu ảnh, người dùng cung cấp cho hệ thống tra cứu ảnh một ảnh mẫu hoặc hình vẽ phác thảo. Sau đó, hệ thống thay đổi các ảnh mẫu này thành biểu diễn trong của các véc tơ đặc trưng. Các khoảng cách giữa các véc tơ đặc trưng của ảnh mẫu hoặc vẽ phác thảo truy vấn của các ảnh trong cơ sở dữ liệu được tính toán và tra cứu được thực hiện với sự hỗ trợ của lược đồ lập chỉ mục. Lược đồ lập chỉ mục cung cấp một cách hiệu quả để tra cứu cơ sở dữ liệu ảnh. Các hệ thống tra cứu gần đây đã kết hợp phản hồi liên quan của người dùng để điều chỉnh quy trình tra cứu nhằm tạo ra các kết quả tra cứu có ý nghĩa hơn về mặt nhận thức và ngữ nghĩa.



Hình I.1. Sơ đồ tra cứu ảnh dựa vào nội dung

Sơ đồ trên Hình I.1 mô tả cơ chế hoạt động của hệ thống tra cứu ảnh dựa vào nội dung. Trong sơ sồ này, cơ sở dữ liệu ảnh sử dụng thủ tục trích rút đặc trưng của các ảnh cơ sở dữ liệu để được một cơ sở dữ liệu đặc trưng. Tương tự, cũng thủ tục trích rút đặc trưng này được sử dụng để trích rút đặc trưng của ảnh truy vấn để được véc tơ đặc trưng của ảnh truy vấn. Tiếp theo, véc tơ đặc trưng của ảnh truy vấn và véc tơ đặc trưng của ảnh cơ sở dữ liệu được so sánh thông qua một độ đo tương tự (hay khoảng cách) nào đó, những ảnh có khoảng cách gần nhau nhất (có độ tương tự cao nhất) được đứng ở đầu của danh sách kết

quả trả về. Trong sơ đồ trên, cơ chế lập chỉ mục được sử dụng để tăng tốc hệ thống tra cứu ảnh, còn cơ chế phản hồi liên quan được sử dụng nhằm chọn những bức ảnh phù hợp với mong muốn của người dùng để nâng cao độ chính xác tra cứu ảnh.

1.2. Mô tả nội dung ảnh

Nhìn chung, nội dung ảnh có thể bao gồm nội dung trực quan của ảnh và nội dung ngữ nghĩa của ảnh. Nội dung trực quan có thể rất chung hoặc theo miền cụ thể. Nội dung ảnh chung bao gồm màu sắc, kết cấu, hình dạng, và quan hệ không gian , ... Nội dung ảnh theo miền cụ thể, như mặt người, phụ thuộc vào ứng dụng và có thể liên quan đến tri thức miền. Nội dung ngữ nghĩa có được bằng cách chú thích ảnh hoặc bằng các thủ tục suy luận phức tạp dựa trên nội dung trực quan.

Một bộ mô tả nội dung trực quan tốt phải bất biến với các thay đổi trong quá trình thu nhận ảnh (ví dụ: sự thay đổi của ánh sáng trong quá trình thu ảnh). Tuy nhiên, có sự cân bằng giữa tính bất biến và năng lực phân biệt của các hình ảnh, vì một lớp bất biến rất rộng làm mất khả năng phân biệt giữa những khác biệt cơ bản. Mô tả bất biến phần lớn đã được nghiên cứu trong thị giác máy tính (như nhận dạng đối tượng), nhưng tương đối mới trong nghiên cứu về tra cứu ảnh.

Việc phân vùng đơn giản không tạo ra các vùng có ý nghĩa về mặt cảm nhận mà là một cách thể hiện chung của ảnh ở độ phân giải tốt hơn. Một phương pháp tốt hơn là chia ảnh thành các vùng đồng nhất theo một số tiêu chí bằng cách sử dụng các thuật toán phân vùng đã được nghiên cứu rộng rãi trong thị giác máy tính. Một cách phức tạp hơn để phân chia một hình ảnh, là thực hiện phân đoạn đối tượng hoàn chỉnh để thu được các đối tượng có ý nghĩa về mặt ngữ nghĩa (như quả bóng, ô tô, con ngựa).

Trong phần này, luận án sẽ giới thiệu một số kỹ thuật được sử dụng rộng rãi để trích rút màu sắc, kết cấu, hình dạng và quan hệ không gian từ hình ảnh. Một số kỹ thuật phân hạng trong tra cứu ảnh dựa vào nội dung [25].

1.2.1. Mầu sắc

Mầu sắc là nội dung trực quan được sử dụng rộng rãi nhất trong tra cứu ảnh. Các giá trị ba chiều của ảnh làm cho khả năng phân biệt của nó vượt trội ảnh đa cấp xám. Trước khi chọn một mô tả màu thích hợp, trước tiên phải xác định không gian màu.

1.2.2. Không gian mầu

Mỗi điểm ảnh (pixel) có thể được biểu diễn dưới dạng một điểm trong không gian màu 3D. Không gian màu thường được sử dụng để tra cứu ảnh bao gồm RGB, Munsell, CIEL*a*b*CIEL*u*v, HSV (hoặc HSL, HSB) và không gian màu ngược nhau. Không có kết hợp nào là tốt nhất, tuy nhiên, một trong những mong muốn của một không gian mầu thích hợp để tra cứu ảnh là tính đồng nhất của nó. Đồng nhất có nghĩa là hai cặp màu có khoảng cách giống nhau trong một không gian màu được xem cảm nhận là như nhau. Nói cách khác, khoảng cách giữa các màu phải liên quan trực tiếp đến sự tương tự giữa chúng.

Không gian RGB là không gian màu được sử dụng rộng rãi để hiển thị hình ảnh. Nó bao gồm ba thành phần là mầu đỏ, mầu xanh lá cây và mầu xanh lam. Các thành phần này gọi là nguyên tố cộng tính vì một mầu trong không gian RGB được tạo ra bằng cách trộn lẫn chúng lại với nhau. Ngược lại, không gian CMY là không gian mầu chủ yếu dùng để in gồm lục lam, đỏ tươi và vàng. Ba thành phần này được gọi là "màu trừ cơ bản" vì một màu trong không gian CMY được tạo ra thông qua sự hấp thụ ánh sáng. Cả không gian RGB và CMY đều phụ thuộc vào thiết bị và không đồng nhất về mặt cảm quan.

Không gian CIEL*a*b* và CIEL*u*v*độc lập với thiết bị và được xem là đồng nhất về mặt cảm nhận. Chúng bao gồm một thành phần độ chói

hoặc độ sáng (L) và hai thành phần mầu a và b hoặc u và v. CIEL*a*b* được thiết kế để xử lý các hỗn hợp chất tạo màu trừ, trong khi CIEL*u*v* được thiết kế để đối phó với các hỗn hợp chất tạo màu cộng tính.

Không gian màu HSV (hoặc HSL, hoăc HSB) được sử dụng rộng rãi trong đồ họa máy tính và là một cách trực quan hơn để mô tả màu sắc, bao gồm ba thành phần là màu sắc, độ bão hòa (độ sáng) và giá trị (độ sáng). Màu sắc luôn thay đổi theo những thay đổi về độ chiếu sáng và hướng camera, do đó nó phù hợp với việc thu hồi đối tượng. Tọa độ RGB có thể dễ dàng được chuyển sang tọa độ HSV (hoặc HLS, hoặc HSB) bởi một công thức đơn giản.

Không gian màu đối lập sử dụng các trục màu đối lập (R-G, 2B-R-G, R+G+B). Biểu diễn này có ưu điểm là cô lập thông tin độ sáng trên trục thứ ba. Với giải pháp này, hai trục màu đầu tiên bất biến với những thay đổi về cường độ chiếu sáng và bóng đổ.

Trong các phần sau, luận án sẽ giới thiệu một số mô tả màu thường được sử dụng gồm: biểu đồ màu, véc tơ liên kết màu, biểu đồ tương quan màu và mô men màu.

1.2.3. Mô men mầu

Mô men mầu đã được sử dụng thành công trong hệ thống tra cứu ảnh [1 , 2] (như QBIC), đặc biệt là khi ảnh chỉ chứa đối tượng. Mô men màu bậc nhất (trung bình), bậc hai (phương sai) và bậc 3 (độ lệch) đã được chứng minh là có hiệu quả và hiệu quả trong việc biểu diễn phân bố màu của ảnh. Về toán học, ba mô men đầu tiên được định nghĩa là:

1.2.4 Biểu đồ mầu

Biểu đồ mầu là sự biểu diễn hiệu quả nội dung màu của ảnh nếu mầu là duy nhất so với phần còn lại của tập dữ liệu. Biểu đồ màu dễ tính toán và hiệu quả trong việc mô tả đặc điểm của sự phân bố màu toàn cục và cục bộ trong

một hình ảnh. Ngoài ra, nó ít bị ảnh hưởng bởi sự dịch chuyển, xoay. và chỉ thay đổi từ từ theo tỉ lệ và góc nhìn.

Vì bất kỳ pixel nào trong ảnh có thể được mô tả bởi ba thành phần trong một không gian màu nhất định (ví dụ: các thành phần đỏ, lục lam trong không gian RGB hoặc màu sắc, độ bão hòa và giá trị trong không gian HSV), một biểu đồ, tức là, phân phối số lượng pixel cho mỗi thùng màu (bin) được lượng hóa, có thể xác được cho từng thành phần. Rõ ràng, biểu đồ màu càng chứa nhiều thùng màu thì nó càng có nhiều khả năng phân biệt. Tuy nhiên, một biểu đồ với số lượng lớn các thùng màu sẽ không chỉ làm tăng chi phí tính toán mà còn không phù hợp để xây dựng các sơ đồ chỉ mục hiệu quả cho cơ sở dữ liệu hình ảnh.

Hơn nữa, lượng tử hóa thùng màu tốt không nhất thiết phải cải thiện hiệu năng tra cứu trong nhiều ứng dụng. Một cách để giảm số lượng thùng là sử dụng không gian màu của trừ để giảm độ sáng của biểu đồ lấy mẫu. Một cách khác là sử dụng phương pháp phân cụm để xác định K màu tốt nhất trong một không gian nhất định cho một tập hợp ảnh nhất định. Mỗi màu tốt nhất này sẽ được lấy làm một thùng của biểu đồ. Vì quá trình phân cụm đó có xem xét đến sự phân bố màu sắc của ảnh trên toàn bộ cơ sở dữ liệu, nên khả năng các thùng màu của biểu đồ không có hoặc rất ít pixel bị rơi sẽ được giảm thiểu. Một tùy chọn khác là sử dụng các thùng màu có số lượng pixel lớn nhất vì một số lượng nhỏ các thùng màu của biểu đồ chiếm phần lớn pixel của ảnh. Việc giảm như vậy không làm giảm hiệu năng của đối sánh biểu đồ mà thậm chí có thể nâng cao nó vì loại đi các thùng màu nhỏ của biểu đồ bị nhiễu.

Khi cơ sở dữ liệu ảnh chứa một số lượng lớn hình ảnh, sự phân biệt so sánh biểu đồ sẽ bão hòa. Để giải quyết vấn đề này, kỹ thuật đối sánh biểu đồ được giới thiệu [10]. Ngoài ra, biểu đồ màu không xem xét thông tin không gian của các pixel, do đó các ảnh rất khác nhau có thể có sự phân bố màu sắc

tương tự. Để tăng khả năng phân biệt, một số cải tiến đã được đề xuất để kết hợp thông tin không gian. Một cách tiếp cận đơn giản là chia ảnh thành các vùng con và tính toán biểu đồ cho từng vùng con đó. Như đã giới thiệu ở trên, việc phân chia có thể đơn giản như một phân vùng hình chữ nhật, hoặc phức tạp như một vùng hoặc thậm chí là phân đoạn đối tượng [26, 27, 28, 29]. Việc tăng số lượng các vùng con làm tăng thông tin về vị trí nhưng cũng làm tăng bộ nhớ và thời gian tính toán.

Do thông tin không gian bổ sung của nó, các véc tơ liên kết màu cung cấp kết quả tra cứu tốt hơn biểu đồ mầu, đặc biệt đối với những ảnh có màu đồng nhất. Ngoài ra, đối với cả biểu đồ màu và biểu diễn véc tơ liên kết màu, không gian màu HSV cung cấp kết quả tốt hơn không gian CIEL*u*v* và CIEL*a*b*.

1.2.5. Biểu đồ mầu tương quan

Biểu đồ tương quan mầu được đề xuất để mô tả không chỉ sự phân bố màu sắc của các pixel, mà còn cả mối tương quan trong không gian của các cặp màu [17]. Chiều thứ nhất và thứ hai của biểu đồ ba chiều là màu của bất kỳ cặp pixel nào và chiều thứ ba là khoảng cách không gian của chúng. Biểu đồ mầu tương quan là một bảng được lập chỉ mục bởi các cặp màu, trong đó mục thứ *k* cho *(i, j)* chỉ định xác suất tìm thấy một pixel màu thứ *j* ở khoảng cách *k* so với một pixel màu *i* trong ảnh. Sau đó, biểu đồ tương quan mầu được định nghĩa là:

$$\gamma_{i,j}^{(k)} = \Pr_{p_1 \in I_{c(i)}, p_2 \in I} \left| p_2 \in I_{c(j)} \right| \left| p_1 - p_2 = k \right| \qquad (1.4)$$

Trong đó i, j ∈ {1, 2, …, N}, k∈ {1, 2, …, d}, và $|p_1 - p_2|$ là khoảng cách giữa các pixel p_1 và p_2. Nếu xem xét tất cả các kết hợp có thể có của các cặp màu thì kích thước của tương quan mầu sẽ rất lớn *($O(N^{2d})$),* do đó, một phiên bản đơn giản của nó được gọi là biểu đồ tự tương quan màu thường được sử dụng thay thế. Phiên bản này làm giảm kích thước xuống *$O(N^d)$*.

1.2.6. Đặc trưng màu

Màu sắc không chỉ phản chiếu chất liệu bề mặt mà còn thay đổi đáng kể theo sự thay đổi của độ chiếu sáng, hướng của bề mặt và hình dạng quan sát của máy ảnh [30, 31]. Sự thay đổi này phải được tính đến. Tuy nhiên, sự bất biến đối với các yếu tố môi trường này không được xem xét trong hầu hết các màu sắc được giới thiệu ở trên.

Gần đây, biểu diễn bất biến màu đã được giới thiệu trong tra cứu ảnh dựa trên nội dung. Một tập hợp các bất biến màu cho tra cứu đối tượng được suy diễn dựa trên mô hình phản xạ đối tượng của Schafer. Biểu diễn bất biến phản xạ, hình dạng và độ chiếu sáng dựa trên véc tơ tỉ lệ xanh lam (r/b, g/b, 1) được đưa ra. Trong [31], đặc trưng bất biến hình học bề mặt được cung cấp.

Mô men màu bất biến này được áp dụng để tra cứu ảnh, có thể mang lại khả năng chiếu sáng, và biểu diễn hình học độc lập với nội dung màu của hình ảnh, nhưng cũng có thể dẫn đến mất một số khả năng phân biệt giữa các hình ảnh.

1.2.7. Đặc trưng kết cấu

Kết cấu là một thuộc tính quan trọng khác của hình ảnh [32]. Các biểu diễn kết cấu khác nhau đã được nghiên cứu trong nhận dạng mẫu và thị giác máy tính. Về cơ bản, các phương pháp biểu diễn kết cấu có thể được phân thành hai loại: cấu trúc và thống kê. Các phương pháp cấu trúc, bao gồm toán tử hình thái biểu đồ kề, mô tả kết cấu bằng cách xác định các cấu trúc nguyên thủy và các quy tắc sắp xếp của chúng, Chúng có xu hướng hiệu quả nhất khi được áp dụng cho các kết cấu đều. Các phương pháp thống kê, bao gồm phổ công suất Fourier, ma trận đồng xuất hiện, phân tích thành phần chính bất biến (SPCA), đặc trưng Tamura, phân rã Wold, trường ngẫu nghiên Markov, mô hình fractal và các kỹ thuật lọc đa phân giải như biến đổi Gabor và Wavelet, đặc trưng kết cấu bởi sự phân bố thống kê của cường độ hình ảnh. đã được sử dụng thường xuyên và đã được chứng minh là có hiệu quả trong việc tra cứu ảnh dựa trên

nội dung.

1.2.8. Đặc trưng Tamura

Tamura bao gồm độ thô, độ tương phản, tính định hướng, độ đều và độ nhám, được thiết kế phù hợp với các nghiên cứu tâm lý về nhận thức của con người về kết cấu [32]. Ba thành phần đầu tiên của Tamura đã được sử dụng trong một số hệ thống tra cứu ảnh nổi tiếng ban đầu, chẳng hạn như QBIC [1, 2] và Photobook.

1.2.9. Độ thô

Độ thô là thước đo độ chi tiết của kết cấu. Để tính toán độ thô, di chuyển trung bình $A_k(x , y)$ được tính trước bằng cách sử dụng cửa số kích thước $2^k \times 2^k$ $(k=0, 1 , \dots , 5)$ tại mỗi pixel là:

$$A_k(x , y) = \sum_{i=x+2^{k-1}}^{x+2^{k-1}-1} . \sum_{j=y+2^{k-1}}^{y+2^{k-1}-1} g(i , j)/2^{2k} \qquad (1. 1)$$

Trong đó $g(i, j)$ là cường độ pixel tại (i, j)

Sau đó, sự khác biệt giữa các cặp đường trung bình cộng không chồng chéo theo hướng ngang và dọc cho mỗi pixel được tính toán, tức là:

$$E_{k ,h}(x , y) = \left|A_k(x + 2^{k-1} , y) - A_k(x - 2^{k-1} , y)\right|$$
$$E_{k ,v}(x , y) = \left|A_k(x , y + 2^{k-1}) - A_k(x , y - 2^{k-1})\right| \qquad (1. 2)$$

Sau đó, giá trị của k tối đa hóa E theo một trong hai hướng được sử dụng để đặt kích thước tốt nhất cho mỗi pixel, tức là:

$$S_{best}(x , y) = 2^k \qquad (1.3)$$

Độ thô sau đó được tính bằng cách lấy S_{best} trung bình trên toàn bộ hình ảnh, tức là:

$$F_{crs} = \frac{1}{m \times n} \sum_{i=1}^{m}. \sum_{j=1}^{n} S_{best}\,(i\,,j) \qquad (1.4)$$

Thay vì lấy giá trị trung bình của S_{best}, có thể thu được phiên bản cải tiến của đặc điểm thô hơn bằng cách sử dụng biểu đồ để mô tả sự phân bố của S_{best}. So với việc sử dụng một giá trị duy nhất để biểu diễn độ thô, việc sử dụng biểu diễn độ thô dựa trên biểu đồ có thể làm tăng đáng kể hiệu năng tra cứu. Sự điều chỉnh này làm cho đặc trưng có khả năng xử lý ảnh hoặc khu vực có nhiều thuộc tính kết cấu và do đó hữu ích hơn cho các ứng dụng tra cứu hình ảnh.

1.2.10. Độ tương phản

Công thức đo độ tương phản như sau:

$$F_{con} = \frac{\sigma}{\propto_4^{1/4}} \qquad (1.5)$$

Trong đó, Kurtosis là thời điểm thứ tư về giá trị trung bình và là phương sai. Công thức này có thể được sử dụng cho cả toàn bộ ảnh và một vùng của ảnh và một vùng của hình ảnh. Với hai vùng 3x3

$\begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$ và $\begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$ và một véc tơ gradient tại mỗi pixel được tính toán.

$$|\Delta G| = (|\Delta_H| + |\Delta_V|)/2$$

$$\theta = tan^{-1}(\Delta_V/\Delta_H) + \pi/2 \qquad (1.6)$$

Trong đó Δ_H và Δ_V là hiệu số theo chiều ngang và chiều dọc của tích chập. Sau đó, bằng cách lượng tử hóa θ và đếm các pixel có độ lớn tương ứng $|\Delta G|$ lớn hơn ngưỡng, có thể xây dựng biểu đồ của θ, ký hiệu là H_D. Biểu đồ này sẽ thể hiện các đỉnh mạnh đối với ảnh có hướng cao và tương đối phẳng đối với ảnh không có hướng mạnh. Sau đó, toàn bộ biểu đồ được tóm tắt để có được một thước đo hướng tổng thể dựa trên độ sắc nét của các đỉnh:

$$F_{dir} = \sum_{P}^{n_P} \sum_{\emptyset \in w_p} \left(\emptyset - H_D(\emptyset)\emptyset_p\right)^2 \quad (1.7)$$

Trong tổng này, p *là* khoảng trên n_p đỉnh; và đối với mỗi đỉnh p, w_p là tập hợp các thùng màu được phân phối trên nó; trong đó $\emptyset_p$ là thùng màu nhận giá trị cao nhất.

1.2.11. Mô hình tự hồi quy đồng thời

Mô hình SAR [33] là một ví dụ của mô hình trường ngẫu nhiên Markov (MRF) [32]. Mô hình này đã rất thành công trong mô hình kết cấu trong những thập kỷ qua. So với các mô hình MRF khác, SAR sử dụng ít tham số hơn. Trong mô hình SAR, cường độ pixel được gọi là biến ngẫu nhiên. Cường độ *g(x , y)* tại pixel *(x , y)* có thể được ước tính dưới dạng kết hợp tuyến tính của các giá trị pixel lân cận *(x' , y')* và số hạng nhiễu cộng *ε(x , y),* tức là:

$$g(x, y) = \mu + \sum_{(x', y') \in D} \theta(x', y') g(x', y') + \varepsilon(x, y) \quad (1.12)$$

Trong đó, μ là giá trị độ lệch được xác định bởi giá trị trung bình của toàn bộ ảnh; D là tập lân cận của *(x, y); θ (x', y')* là một tập hợp các trọng số được liên kết với mỗi pixel lân cận; *ε(x, y)* là một biến ngẫu nhiên Gaussian độc lập với trung bình bằng không và phương sai bằng σ^2.

Các thông số θ và σ được sử dụng để đo kết cấu. Ví dụ: giá trị σ cao hơn có nghĩa là độ chi tiết tốt hơn hoặc ít thô hơn; giá trị *θ(x, y+1) và θ(x, y-1)* cao hơn cho biết kết cấu được định hướng theo chiều dọc. Kỹ thuật sai số bình phương nhỏ nhất (LSE) hoặc phương pháp ước lượng khả năng xảy ra tối đa (MSE) thường được sử dụng để ước tính các tham số của mô hình SAR. Mô hình SAR không phải là bất biến quay. Để thu được mô hình SAR bất biến xoay (RISAR) [34, 35], các pixel nằm trên các vòng tròn có bán kính khác nhau

được căn giữa tại mỗi pixel (x, y) đóng vai trò là tập lân cận của nó. Do đó, cường độ g(x , y) tại pixel (x , y) có thể được ước tính là:

$$g(x , y) = \mu + \sum_{i=1}^{p} \theta_i(x , y) l_i(x , y) + \varepsilon(x , y) \quad (1.13)$$

Với p là số lân cận hình tròn. Để giảm chi phí tính toán và đồng thời đạt được sự bất biến quay, p không được quá lớn cũng không được quá nhỏ. Thường $p=2.l(x, y)$ có thể được tính bằng:

$$l_i(x , y) = \frac{1}{8i} + \sum_{(x' ,y')\in N_i} w_i (x' , y') g(x' , y') \quad (1.14)$$

Với N_i là lân cận hình tròn thứ i của (x, y); $w_i(x', y')$ là một tập hợp các trọng số được tính toán trước cho biết sự đóng góp của pixel (x', y') trong vòng tròn thứ i.

Để mô tả các kết cấu có độ chi tiết khác nhau, mô hình hồi quy tự đồng thời đa độ phân giải (MRSAR) đã được đề xuất để cho phép phân tích kết cấu đa tỷ lệ. Ảnh được biểu diễn bằng kim tự tháp Gaussian đa độ phân giải với bộ lọc thông thấp và lấy mẫu phụ được áp dụng ở một số cấp độ liên tiếp. Sau đó, mô hình SAR hoặc RISAR có thể được áp dụng cho mỗi cấp kim tự tháp.

MRSAR đã được chứng minh có hiệu năng tốt hơn trên cơ sở dữ liệu kết cấu Brodatz so với nhiều đặc trưng kết cấu khác, chẳng hạn như phân tích thành phần chính, phân hủy Wold và đã biến đổi wavelet.

1.2.12. Bộ lọc Gabor

Bộ lọc Gabor đã được sử dụng rộng rãi để trích rút đặc trưng của hình ảnh, đặc biệt là các đặc trưng về kết cấu [32]. Nó được sử dụng để phát hiện cạnh, đường nét và các đặc trưng hình học trong hình ảnh và thường được sử dụng như một máy dò cạnh, vạch định hướng, và điều chỉnh tỉ lệ. Đã có nhiều cách tiếp cận được đề xuất để mô tả các kết cấu của ảnh dựa trên bộ lọc Gabor. Ý

tưởng cơ bản của việc sử dụng bộ lọc Gabor để trích rút kết cấu như sau: Hàm Gabor hai chiều g(x , y) được định nghĩa là:

$$G(x , y) = \frac{1}{2\pi\sigma_x\sigma_y} exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + 2\pi jWx\right] \quad (1.5)$$

Trong đó, σ_x *và* σ_y là độ lệch chuẩn của đường bao Gaussian dọc theo hướng x và y. Sau đó, một bộ lọc Gabor có thể thu được bằng co dãn và xoay thích hợp $g(x , y)$:

$$g_{mn}(x , y) = a^{-m} g(x' , y') \quad (1.16)$$
$$x' = a^{-m}(x\cos\theta + y\sin\theta)$$
$$y' = a^{-m}(-x\sin\theta + y\cos\theta)$$

Trong đó, $a > 1$, $\theta = n\pi/K$, $n = 0, 1, \ldots, K-1$ và $m = 0, 1, \ldots, S-1$. K và S là số định hướng và tỷ lệ. Hệ số tỷ lệ đảm bảo rằng năng lượng không phụ thuộc vào m. Cho một ảnh $I(x , y)$, phép biến đổi Gabor của nó được định nghĩa là:

$$W_{mn}(x , y) = \int I(x , y) g^*_{mn}(x - x_1 , y - y_1) dx_1 dy_1 \quad (1.17)$$

Trong đó * chỉ ra liên hợp phức tạp. Sau đó, giá trị trung bình và độ lệch chuẩn của độ lớn $W_{mn}(x , y)$ là $f = [\mu_{00}, \sigma_{00}, \ldots, \mu_{mn}, \sigma_{mn}, , \mu_{S-1k-1}, \sigma_{S-1K-1}]$

Có thể sử dụng để thể hiện kết cấu của một vùng kết cấu đồng nhất.

1.2.13. Biến đổi Wavelet

Tương tự như lọc Gabor, biến đổi Wavelet cung cấp một cách tiếp cận đa độ phân giải để phân tích kết cấu.

$$\psi_{mn}(x) = 2^{-m/2}\psi(2^{-m}x - n) \quad (1.18)$$

Trong đó m và n là các tham số giãn dịch. Tín hiệu $f(x)$ có thể được biểu diễn dưới dạng:

$$f(x) = \sum_{m,n} c_{mn}\psi_{mn}(x) \quad (1.19)$$

Việc tính toán các phép biến đổi Wavelet của tín hiệu 2D liên quan đến việc lọc đệ quy và lấy mẫu con. Ở mỗi mức, tín hiệu được phân tách thành các

dải tần con là LL, LH và HH, trong đó L biểu thị tần số thấp và H biểu thị tần số cao. Hai loại biến đổi Wavelet chính được sử dụng để phân tích kết cấu là biến đổi Wavelet có cấu trúc kim tự tháp (PWT) và biến đổi Wavelet có cấu trúc cây (TWT). PWT phân rã đệ quy dải LL. Tuy nhiên, đối với một số kết cấu, thông tin quan trọng nhất thường xuất hiện trong các kênh tần số trung bình. Để khắc phục nhược điểm này, TWT sẽ phân hủy các băng khác như LH, HL hoặc HH khi cần thiết.

Sau khi phân rã, các véc tơ, các véc tơ đặc trưng có thể được xây dựng bằng cách sử dụng giá trị trung bình và độ lệch chuẩn của sự phân bố năng lượng của mỗi dải con ở mỗi mức. Đối với phân rã ba cấp, PWT dẫn đến một véc tơ đặc trưng của các thành phần 3x4x2. Đối với TWT, đặc trưng sẽ phụ thuộc vào phân rã các dải con ở mỗi cấp. Cây phân hủy cố định có thể thu được bằng cách phân hủy tuần tự các dải LL, LH và HL, và do đó dẫn đến véc tơ đặc trưng của các thành phần 52x2. Lưu ý rằng trong ví dụ này, đối tượng địa lý do PWT thu được có thể được coi là tập hợp con của đối tượng địa lý do TWT thu được. Hơn nữa, theo sự so sánh của biến đổi Wavelet khác nhau, lựa chọn cụ thể của bộ lọc Wavelet không quan trọng đối với phân tích kết cấu.

1.2.14. Đặc trưng hình dạng

Hình dạng của các đối tượng hoặc vùng đã được sử dụng trong nhiều hệ thống tra cứu ảnh dựa trên nội dung [2, 36, 37, 38, 39]. So với các hình dạng và kết cấu, các hình dạng thường được mô tả sau khi ảnh đã được phân đoạn thành các vùng hoặc đối tượng. Vì khó đạt được sự phân đoạn ảnh mạnh và chính xác, việc sử dụng hình dạng để tra cứu ảnh đã bị giới hạn trong các ứng dụng chuyên biệt mà gồm các đối tượng hoặc khu vực có sẵn. Các phương pháp hiện đại để mô tả hình dạng có thể được phân loại thành dựa trên ranh giới (hình dạng tuyến tính, xấp xỉ đa giác, mô hình phần tử hữu hạn và mô tả hình dạng dựa trên Fourier) hoặc các phương pháp dựa trên vùng (mô men thống

kê). Một đặc trưng biểu diễn hình dạng tốt cho một đối tượng phải bất biến đối với phép dịch chuyển, xoay và chia tỉ lệ. Trong phần này, luận án mô tả ngắn gọn một số hình dạng này thường được sử dụng trong các ứng dụng tra cứu hình ảnh. Để có cái nhìn tổng quan ngắn gọn về các kỹ thuật kết hợp hình dạng.

1.2.15. Mô men bất biến

Mô men bất biến được gọi là 'invariant moment' [40] là tập hợp các đặc trưng số học của hình ảnh được tính toán dựa trên các giá trị cường độ của điểm ảnh trong hình ảnh. Mục đích của việc sử dụng mô men bất biến là để tạo ra các đặc trưng có tính chất không thay đổi khi ảnh bị thay đổi bởi các biến đổi hình học như quay, phóng to, thu nhỏ hoặc lật đối xứng, điều này giúp cho việc nhận dạng và phân loại đối tượng trở nên ổn định hơn ttrong các tình huống khác nhau.

1.2.16. Góc quay

Góc quay thể hiện mức độ xoay của hình ảnh quanh một trục tương ứng. Trong không gian hai chiều, góc quay được đo bằng độ và thường được tính theo chiều kim đồng hồ. Trong xử lý ảnh, để biến đổi xoay thường sử dụng biến đổi hình học như ma trận xoay. Ma trận xoay 2x2 và góc quay được tính theo radian. Ma trận xoay áp dụng lên các điểm ảnh trong hình ảnh để thực hiện biến đổi xoay. Biến đổi xoay sử dụng trong việc tạo ra các phiên bản xoay của ảnh để tạo ra dữ liệu đào tạo đa dạng hơn trong mô hình học máy.

1.2.17. Mô tả Fourier

Biến Fourier dựa trên ý tưởng mọi tín hiệu (bao gồm cả hình ảnh) có thể được biểu diễn bằng cách kết hợp giữa sóng sin và cos có tần số và biên độ khác nhau [41]. Biến Fourier giúp chuyển từ miền thời gian sang miền tần số, từ đó làm cho việc phân tích và xử lý tín hiệu trở lên thuận tiện hơn.

Trong xử lý ảnh biến Fourier thường được sử dụng để phân tích tần số, loại bỏ nhiễu, nén ảnh.

Phân tích tần số : Biến Fourier cho phép phân tích một hình ảnh thành các tần số khác nhau. Các thành phần tần số này thể hiện các mẫu sóng trong hình ảnh và cho biết các tần số khác nhau đang xuất hiện trong ảnh.

Loại bỏ nhiễu : Bằng cách chuyển hình ảnh sang miền tần số và loại bỏ các thành phần tần số thấp (đại diện cho nhiễu) để làm sạch ảnh và giảm thiểu nhiễu.

Nén ảnh : Biến Fourier cho phép nén ảnh bằng cách chỉ giữ lại các thành phần tần số quan trọng, từ đó giảm dung lượng của ảnh.

Xử lý và cải thiện hình ảnh : Bằng cách thay đổi các thành phần tần số hoặc áp dụng biến đổi ngược ta có thể thay đổi hình dạng và tính chất của ảnh. Tóm lại : biến Fourier là một công cụ tốt trong xử lý ảnh giúp phân tích và xử lý tín hiệu ảnh dựa trên phổ tần số của chúng.

1.2.18. Tính tuần hoàn, độ lệch tâm và hướng trục chính

Tính tuần hoàn được tính là :

$$\alpha = \frac{4\pi S}{P^2} \qquad (1.20)$$

Trong đó, S là kích thước và P là chu vi của một vật thể. Giá trị này tương ứng với một đường tròn hoàn hảo.

Hướng trục chính có thể được xác định là hướng của ký hiệu riêng lớn nhất của ma trận hiệp phương sai bậc hai của một vùng hoặc một đối tượng. Độ lệch tâm có thể được định nghĩa là tỷ số giữa giá trị riêng nhỏ nhất và giá trị riêng lớn nhất.

1.2.19. Thông tin không gian

Có thể dễ dàng phân biệt các vùng hoặc đối tượng có màu sắc và kết cấu tương tự bằng cách áp đặt các ràng buộc về không gian [41]. Ví dụ : các vùng của bầu trời xanh và đại dương có thể có biểu đồ màu tương tự nhau, nhưng vị trí không gian của chúng trong ảnh là rất khác nhau. Do đó, vị trí không gian của các vùng (hoặc đối tượng) hoặc mối quan hệ không gian giữa nhiều vùng

(hoặc đối tượng) hoặc mối quan hệ không gian giữa nhiều vùng (hoặc đối tượng) trong ảnh rất hữu ích cho việc tra cứu ảnh.

Biểu diễn mối quan hệ không gian được sử dụng rộng rãi nhất là các chuỗi 2D do Chang và cộng sự đề xuất. Nó được xây dựng bằng cách chiếu các ảnh dọc theo các hướng x và y. Hai bộ ký hiệu, V và A, được xác định trên hình chiếu. Mỗi ký hiệu trong V đại diện cho một đối tượng trong ảnh. Mỗi ký hiệu trong A đại diện cho một kiểu quan hệ không gian giữa các đối tượng. Là biến thể của nó, 2D G-string cắt tất cả các đối tượng dọc theo hộp giới hạn tối thiểu của chúng và mở rộng các mối quan hệ không gian thành hai tập hợp các toán tử không gian. Phương thức còn lại xác định các mối quan hệ không gian toàn cục, chỉ ra rằng hình chiếu của hai đối tượng là rời nhau, liền kề hoặc nằm ở cùng một vị trí.

Ngoài ra, 2D C-string được đề xuất để giảm thiểu số lượng đối tượng cắt. Chuỗi 2D-B đại diện cho một đối tượng bằng hai ký hiệu, đại diện cho ranh giới đầu và cuối của đối tượng. Tất cả các phương thức này có thể hỗ trợ ba loại truy vấn. Truy vấn loại 0 tìm tất cả các ảnh có chứa đối tượng $O_1, O_2, \ldots, O_n$. Loại 1 tìm tất cả các ảnh chứa các đối tượng có mối quan hệ nhất định với nhau, nhưng khoảng cách giữa chúng là không đáng kể. Loại 2 tìm tất cả các ảnh có quan hệ khoảng cách nhất định với nhau. Ngoài chuỗi 2D, cây tứ phân không gian và ảnh biểu tượng cũng được sử dụng để biểu diễn thông tin không gian. Tuy nhiên, việc tra cứu ảnh dựa trên các mối quan hệ không gian của các vùng vẫn là một bài toán nghiên cứu khó trong tra cứu ảnh dựa trên nội dung, bởi vì việc phân đoạn các đối tượng hoặc vùng một cách đáng tin cậy thường không khả thi ngoại trừ các ứng dụng rất hạn chế. Mặc dù một số hệ thống chỉ đơn giản phân chia các ảnh thành các khối con thông thường, nhưng chỉ đạt được thành công hạn chế với các sơ đồ phân chia không gian như vậy vì hầu

hết các ảnh tự nhiên không bị giới hạn về mặt không gian thành các khối con thông thường.

1.3. Các kỹ thuật tương tự và các lược đồ lập chỉ mục

Thay vì đối sánh chính xác, tra cứu ảnh dựa trên nội dung sẽ tính toán các điểm tương tự trực quan giữa ảnh truy vấn và ảnh trong cơ sở dữ liệu. Theo đó, kết quả tra cứu không phải là một ảnh đơn lẻ mà là một danh sách các ảnh được xếp hàng theo độ tương tự của chúng với ảnh truy vấn. Nhiều kỹ thuật tương tự đã được phát triển cho các ước tính thực nghiệm dựa trên tra cứu ảnh về sự phân bố của các đối tượng trong những năm gần đây. Các kỹ thuật tương tự và khoảng cách khác nhau sẽ ảnh hưởng đáng kể đến hiệu năng tra cứu của hệ thống tra cứu ảnh. Trong phần này, luận án sẽ giới thiệu một số kỹ thuật tương tự thường được sử dụng. Luận án ký hiệu *D(I, J)* là độ đo khoảng cách giữa ảnh truy vấn I và hình ảnh J trong cơ sở dữ liệu, và $f_i(I)$ là số pixel trong bin của ảnh truy vấn I.

1.3.15. Khoảng cách Minkowski

Nếu mỗi chiều của véc tơ đặc trưng của ảnh là độc lập với nhau và có tầm quan trọng như nhau, thì khoảng cách dạng Minkowski L_p thích hợp để tính khoảng cách giữa hai ảnh. Khoảng cách này được xác định là:

$$D(I, J) = \left(\sum_i |f_i(I) - f_i(J)|^P \right)^{\frac{1}{P}} \quad (1.21)$$

Khi $p = 1, 2$ và ∞, *D(I, J)* lần lượt là khoảng cách L_1, L_2 (còn gọi là khoảng cách Euclide) và L_∞. Khoảng cách dạng Minkowski được sử dụng rộng rãi nhất để tra cứu hình ảnh. Ví dụ, hệ thống MARS [42] đã sử dụng khoảng cách Euclide để tính toán sự tương tự giữa các kết cấu; Netra đã sử dụng khoảng cách Euclide cho màu sắc và hình dạng, và khoảng cách L_1 cho họa tiết; Blobworlk đã sử dụng khoảng cách Euclide cho đặc trưng kết cấu và hình dạng.

Ngoài ra, Voorhees và Poggio đã sử dụng khoảng cách L_∞ để tính toán sự tương tự giữa các kết cấu của hình ảnh.

Giao giữa các biểu đồ có thể được coi là một trường hợp đặc biệt của khoảng cách L_1, được sử dụng bởi Swain và Ballard để tính độ tương tự giữa các ảnh màu. Giao của hai biểu đồ I và J được xác định là:

$$S(I , J) = \frac{\sum_{i=1}^{N} \min (f_i(I) , f_i(j))}{\sum_{i=1}^{N} f_i(j)} \quad (1.22)$$

Nó đã được chứng minh rằng giao của hai biểu đồ ít nhạy cảm với những thay đổi về độ phân giải hình ảnh, kích thước biểu đồ, độ kín, độ sâu và điểm xem.

1.3.16. Khoảng cách toàn phương

Khoảng cách Minkowski xử lý tất cả các thùng màu của biểu đồ cho đối tượng hoàn toàn độc lập và không tính đến thực tế là các cặp mầu trùng tương ứng với các đối tượng địa lý giống nhau hơn các cặp khác về mặt cảm quan. Để giải quyết vấn đề này, khoảng cách dạng toàn phương được giới thiệu:

$$D(i, j) = \sqrt{(F_i - F_j)^T A (F_i - F_j)} \quad (1.23)$$

Trong đó, $A=[a_{ij}]$ là ma trận tương tự và a_{ij} biểu thị sự tương tự giữa bin i và j của $f_i(I)$ và $f_i(J)$.

Khoảng cách dạng toàn phương đã được sử dụng trong nhiều hệ thống tra cứu để tra cứu hình ảnh dựa trên biểu đồ màu. Nó được chỉ ra rằng khoảng cách dạng toàn phương có thể dẫn đến kết quả mong muốn về mặt cảm quan hơn so với khoảng cách Euclide và phương pháp giao của hai biểu đồ vì nó xem xét sự giống nhau chéo giữa các màu.

1.3.17. Khoảng cách Mahalanobis

Số liệu khoảng cách Mahalanobis thích hợp khi mỗi chiều của véc tơ đặc trưng hình ảnh phụ thuộc lẫn nhau và có tầm quan trọng khác nhau. Nó được định nghĩa là:

$$D(i,j)= \sqrt{(F_i - F_j)^T C^{-1}(F_i - F_j)} \quad (1.24)$$

Khi đó C là ma trân hiệp phương sai của các véc tơ đặc trưng.

Khoảng cách Mahalanobis có thể được đơn giản hóa nếu các kích thước của đối tượng địa lý là độc lập. Trong trường hợp này, chỉ cần một phương sai của từng thành phần đặc trưng, c_i là cần thiết.

$$D(i,j)= \sum_{i=1}^{N} \frac{(F_i - F_j)^{\wedge 2}}{C_i} \quad (1.25)$$

1.3.18. Phân kỳ Kullback-Leibler và Jeffrey-Divergence

Sự phân kỳ *Kullback-Leibler (KL)* đo lường mức độ khác biệt giữa hai phân phối đặc trưng. Độ phân kì KL giữa hai ảnh I và J được xác định là:

$$D(i,j)= \sum_i f_i(I) \log \frac{f_i(I)}{f_i(J)} \quad (1.26)$$

Sự phân kỳ KL được sử dụng trong [43] làm thước đo độ tương tự cho kết cấu. Sự phân kỳ Jeffrey (JD) xác định bởi:

$$D(i,j)= \sum_i f_i(I) \log \frac{f_i(I)}{f_i} + f_i(j) \log \frac{f_i(j)}{f_i} \quad (1.27)$$

Trong đó, $f_i = [f_i(I) + f_j(j)/2]$. Ngược lại với phân kỳ KL, JD là đối xứng và ổn định hơn về mặt số khi so sánh hai phân phối thực nghiệm.

1.3.19. Lập chỉ mục

Một vấn đề quan trọng khác trong tra cứu hình ảnh dựa trên nội dung là lập chỉ mục hiệu quả và tra cứu nhanh hình ảnh dựa trên trực quan. Bởi vì các véc tơ đặc trưng của hình ảnh có xu hướng có chiều cao và do đó không phù

hợp với cấu trúc lập chỉ mục truyền thống, nên việc giảm chiều thường được sử dụng trước khi thiết lập một lược đồ lập chỉ mục hiệu quả.

Một trong những kỹ thuật thường được sử dụng để giảm chiều là phân tích thành phần chính (PCA). Đây là một kỹ thuật tối ưu ánh xạ tuyến tính dữ liệu đầu vào vào một không gian tọa độ sao cho các trục được căn chỉnh để phản ánh các biến thể tối đa trong dữ liệu. Hệ thống QBIC sử dụng PCA để giảm véc tơ đặc trưng. Ngoài PCA, nhiều nhà nghiên cứu đã sử dụng phép biến đổi *Karhunen-Loeve (KL)* để giảm chiều của không gian đối tượng. Mặc dù phép biến đổi KL có một số hữu ích như khả năng định vị không gian con quan trọng nhất, nhưng để xác định sự tương tự của mẫu có thể bị phá hủy trong quá trình giảm chiều. Ngoài biến đổi PCA và KL, mạng nơ – ron cũng đã được chứng minh là một công cụ hữu ích để giảm chiều của không gian dữ liệu.

Sau khi giảm chiều, dữ liệu nhiều chiều được lập chỉ mục. Một số cách tiếp cận đã được đề xuất cho mục đích này, bao gồm *R-tree* (đặc biệt *R*-tree*), cây tứ phân tuyến tính, cây *K-d-B* và các tệp lưới. Hầu hết các phương pháp lập chỉ mục đa chiều này có hiệu năng hợp lý đối với một số lượng nhỏ chiều (tối đa 20 chiều), nhưng khám phá theo cấp số nhân với sự gia tăng của chiều và cuối cùng giảm xuống tra cứu tuần tự. Hơn nữa, các lược đồ lập chỉ mục này giả định rằng việc so sánh đặc trưng cơ bản dựa trên khoảng cách Euclide, điều này không đúng với nhiều ứng dụng tra cứu hình ảnh. Một nỗ lực để giải quyết các vấn đề về lập chỉ mục là sử dụng lược đồ lập chỉ mục phân cấp dựa trên bản đồ tự tổ chức *(SOM)*. Ngoài việc mang lại lợi ích cho việc lập chỉ mục, *SOM* cung cấp cho người dùng một công cụ hữu ích để duyệt các hình ảnh đại diện của từng loại.

1.4. Tương tác người dùng

Đối với tra cứu hình ảnh dựa trên nội dung, tương tác của người dùng với hệ thống tra cứu là rất quan trọng vì nó có thể sửa đổi linh hoạt các truy vấn

bằng cách để người dùng tham gia vào quá trình tra cứu. Giao diện người dùng trong hệ thống tra cứu hình ảnh bao gồm phần tạo truy vấn và phần trình bày kết quả.

1.4.1. Kỹ thuật truy vấn bởi phác thảo

Truy vấn bằng phác thảo và truy vấn theo mẫu là vẽ một bản phác thảo hoặc cung cấp một hình ảnh mẫu mà từ đó các hình ảnh có các nội dung trực quan tương tự sẽ được trích xuất từ cơ sở dữ liệu.

Truy vấn bằng phác thảo cho phép người dùng vẽ phác thảo hình ảnh bằng công cụ chỉnh sửa đồ họa được cung cấp bởi hệ thống tra cứu hoặc một số phần mềm khác. Các truy vấn có thể được hình thành bằng cách vẽ một số đối tượng có các thuộc tính nhất định như màu sắc, kết cấu, hình dạng, kích thước và vị trí. Trong hầu hết các trường hợp, một bản phác thảo thô là đủ, vì truy vấn có thể được tinh chỉnh dựa trên kết quả truy xuất.

1.4.2. Phản hồi liên quan

Nhận thức của con người về sự tương tự của hình ảnh là chủ quan, ngữ nghĩa và phụ thuộc vào nhiệm vụ. Mặc dù các phương pháp dựa trên nội dung cung cấp các hướng đi đầy hứa hẹn để tra cứu hình ảnh, nhưng nhìn chung, kết quả tra cứu dựa trên sự tương tự của các đặc trưng hình ảnh thuần túy không nhất thiết phải có ý nghĩa về mặt nhận thức và ngữ nghĩa. Ngoài ra, mỗi loại đặc trưng hình ảnh có xu hướng chỉ nắm bắt một khía cạnh của thuộc tính hình ảnh và người dùng thường khó xác định rõ ràng cách kết hợp các khía cạnh khác nhau. Để giải quyết những vấn đề này, phản hồi liên quan tương tác, một kỹ thuật trong hệ thống tra cứu thông tin dựa trên văn bản truyền thống đã được giới thiệu. Với phản hồi về mức độ liên quan, có thể thiết lập mối liên hệ giữa các khái niệm cấp cao và các đối tượng địa lý cấp thấp.

Phản hồi về mức độ liên quan là một kỹ thuật học tích cực có giám sát được sử dụng để nâng cao hiệu quả của hệ thống thông tin. Ý tưởng chính là sử

dụng các mẫu tích cực và tiêu cực từ người dùng để cải thiện hiệu năng hệ thống. Đối với một truy vấn nhất định, trước tiên, hệ thống sẽ tra cứu danh sách các hình ảnh được xếp hạng theo độ đo tương tự được xác định trước. Sau đó, người dùng đánh dấu các hình ảnh được tra cứu là có liên quan (mẫu tích cực) với truy vấn hoặc không kiên quan (mẫu tiêu cực). Hệ thống sẽ tinh chỉnh kết quả tra cứu dựa trên phản hồi và hiển thị danh sách hình ảnh mới cho người dùng. Do đó, vấn đề quan trọng trong phản hồi về liên quan là làm thế nào để kết hợp các mẫu tích cực và tiêu cực để tinh chỉnh truy vấn và/hoặc để điều chỉnh độ đo tương tự.

1.4.3. Đánh giá hiệu năng

Để đánh giá hiệu năng của hệ thống tra cứu, hai độ đo cụ thể là truy hồi (recall) và độ chính xác (precision), được lấy từ tra cứu thông tin truyền thống. Đối với truy vấn q, tập dữ liệu hình ảnh trong cơ sở dữ liệu có liên quan đến truy vấn q được ký hiệu là R(q). Độ chính xác của tra cứu được định nghĩa là phần nhỏ của các hình ảnh được tra cứu thực sự có liên quan đến truy vấn:

$$precision = \frac{|Q(q) \cap R(q)|}{|Q(q)|} \tag{1.28}$$

Phần truy hồi là phần hình ảnh có liên quan được trả về bởi truy vấn:

$$recall = \frac{|Q(q) \cap R(q)|}{|R(q)|} \tag{1.29}$$

Thông thường, cần sự cân bằng giữa hai phương pháp này bởi vì việc cải thiện truy hồi sẽ có thể phải hy sinh độ chính xác. Trong các hệ thống tra cứu điển hình, việc truy hồi có xu hướng tăng lên khi số lượng các mục được tra cứu tăng lên; đồng thời độ chính xác có thể giảm. Ngoài ra, việc chọn tập dữ liệu có liên quan R(q) kém ổn định hơn nhiều do các hình ảnh diễn giải khác nhau. Hơn nữa, khi số lượng hình ảnh có liên quan lớn hơn số lượng hình ảnh

được tra cứu, việc truy hồi là vô nghĩa. Do đó, độ chính xác và truy hồi chỉ là những mô tả sơ bộ về hiệu năng của hệ thống tra cứu.

Gần đây MPEG7 đề xuất một phương pháp đánh giá hiệu năng tra cứu mới, xếp hạng tra cứu đã sửa đổi chuẩn hóa trung bình (ANMRR) [31]. Nó kết hợp độ chính xác và truy hồi để có được một độ đo khách quan nhất. Biểu thị số lượng hình ảnh chân lý cơ bản cho một truy vấn q nhất định là N(q) và số lượng hình ảnh chân lý cơ bản tối đa cho tất cả các truy vấn Q, tức là, *max(N(q1), N(q2), …, N(qQ)), như M.* Sau đó, đối với một truy vấn q cho trước, mỗi hình ảnh chân lý cơ bản k được gán một giá trị thứ hạng (k) tương đương với thứ hạng của nó trong các hình ảnh chân lý tin cậy nếu nó nằm trong K đầu tiên (trong đó K = min[4N(q), 2M]); hoặc giá trị xếp hạng K+1 nếu không. Xếp hạng trung bình AVR(q) cho truy vấn q được tính như sau:

$$AVR(q) = \sum_{k=1}^{N(q)} \frac{rank(k)}{N(q)} \quad (1.30)$$

Xếp hạng tra cứu sửa đổi MRR(q) được tính như sau:

$$MRR(q) = AVR(q) - 0,5 - 0,5 * N(q) \quad (1.31)$$

MRR(q) nhận giá trị 0 khi tất cả các hình ảnh chân thực tin cậy nằm trong K kết quả tra cứu đầu tiên.

Xếp hạng tra cứu đã sửa đổi chuẩn hóa *NMRR(q)* nằm trong khoảng từ 0 đến 1, được tính như sau:

$$NMRR(q) = \frac{MRR(q)}{K+ - 0,5 - 0,5 * N(q)} \quad (1.32)$$

Sau đó, xếp hạng tra cứu được sửa đổi chuẩn hóa trung bình ANMRR trên tất cả các truy vấn Q được tính như sau:

$$ANMRR = \frac{1}{Q} \sum_{q=1}^{Q} NMRR(q) \quad (1.33)$$

1.5. Giảm khoảng cách ngữ nghĩa

1.5.1. Khái niệm

Khoảng cách ngữ nghĩa là một trong những ví dụ điển hình trong tra cứu ảnh dựa vào nội dung. Khoảng cách ngữ nghĩa là khoảng cách đề cập đến mức độ tương đồng hoặc sự giống nhau (khoảng cách) giữa nhận thức của con người và sự hiểu biết có được từ các thuật toán máy tính về cùng một ảnh. Khoảng cách này có ảnh hưởng trực tiếp đến việc đánh giá các ảnh là tương tự bởi các thuật toán. Sự tương tự về ảnh được xác định bởi một người quan sát trong ngữ cảnh cụ thể ở cấp độ ngữ nghĩa cao.

Mặt khác, đối với các thuật toán, độ tương tự của ảnh được xác định bằng các phân tích tổng hợp của các giá trị pixel liên quan đến màu sắc, kết cấu hoặc hình dạng. Khoảng cách ngữ nghĩa được kết nối chặt chẽ với không chỉ nội dung (đối tượng) của ảnh mà còn với các đặc trưng được sử dụng cho dấu hiệu và hiệu quả của các thuật toán được sử dụng để suy ra nội dung hình ảnh. Khoảng cách ngữ nghĩa có tầm quan trọng cao như là một yếu tố ảnh hưởng đến tính hữu dụng của hệ thống CBIR và thường được các nhà nghiên cứu CBIR trích dẫn. Ba mức độ:

Thứ nhất: Bai, J. Chen, L. Huang, K. Kpalma, & S. Chen đã cung cấp phân tích chi tiết về khoảng cách ngữ nghĩa và tạo ra một phân loại các ảnh và kiểu người dùng để hiểu thêm về các loại khoảng cách ngữ nghĩa.

Thứ hai: Eakins và Graham [37] đã sử dụng ý tưởng về nội dung ngữ nghĩa như một cách để phân loại các truy vấn CBIR cụ thể: Eakins định nghĩa ba loại truy vấn CBIR theo mức độ nội dung ngữ nghĩa của chúng.

Thứ ba: Bosch và các cộng sự đã đề xuất một phương pháp nhằm thu hẹp khoảng cách ngữ nghĩa bằng các phương pháp. Trong lĩnh vực ảnh cảnh tự nhiên.

1.5.2. Một số nghiên cứu theo hương tiếp cận học có giám sát

Một số kỹ thuật học có giám sát, như máy véc tơ hỗ trợ (SVM), phân lớp Bayes, thường được đưa vào các hệ thống tra cứu ảnh dựa vào nội dung nhằm mục đích học khái niệm ngữ nghĩa mức cao từ đặc trưng mức thấp.

Với cơ sở lý thuyết vững chắc, SVM sử dụng để nhận dạng đối tượng, phân lớp văn bản… và là một thuật toán học tốt trong hệ thống tra cứu ảnh. Ban đầu, SVM được thiết kế để phân lớp nhị phân. Giả sử, có một tập dữ liệu huấn luyện $\{x_1, x_2, ..., x_n\}$ là các véc tơ trong không gian $X \subseteq R^d$ thuộc hai lớp riêng biệt với tập nhãn $\{y_1, y_2,..., y_n\}$ và y_i $\{-1, 1\}$. Muốn tìm một mặt để tách biệt dữ liệu, mặt phân tách tối ưu (OSP) là một trong những lề cực đại (khoảng cách giữa mặt và điểm dữ liệu của mỗi lớp). Để học đa khái niệm về tra cứu ảnh, một SVM được huấn luyện cho mỗi khái niệm. Một phương pháp được dùng rộng rãi nữa là phân lớp Bayesian.

Sử dụng phân lớp nhị phân Bayesian, khái niệm mức cao về cảnh thiên nhiên thu được từ các đặc trưng mức thấp. Hệ thống phân lớp tự động ảnh cơ sở dữ liệu thành một nhóm như trong nhà/ ngoài trời, và hình ảnh ngoài trời được phân thành thành phố và cảnh quan. Mạng Bayesian được dùng để phân lớp ảnh trong nhà/ ngoài trời. Các kỹ thuật học khác như mạng nơ ron được dùng cho học khái niệm. Đầu tiên, người ta lựa chọn 11 nhóm khái niệm: gạch, mây, lông, cỏ, đá, kem, kính, đường, đá, cát, da, cây và nước. Sau đó, một lượng lớn dữ liệu huấn luyện (đặc trưng mức thấp của các vùng) được đưa vào phân lớp mạng nơ ron để thiết lập liên kết giữa đặc trưng trực quan mức thấp của một ảnh và ngữ nghĩa mức cao của nó (nhãn loại). Một bất lợi của thuật toán này là đòi hỏi một lượng lớn dữ liệu huấn luyện và cần những tính toán phức

tạp. Trong nghiên cứu về độ phức tạp trong số lượng dữ liệu lớn trong huấn luyện có đề cập tới các thuật toán học thường có hai vấn đề: (1) cần một lượng lớn các mẫu huấn luyện có nhãn (2) Tập huấn luyện được cố định trong suốt quá trình học và ứng dụng. Vì thế, nếu ứng dụng thay đổi, các mẫu nhãn mới cần phải cung cấp để đảm bảo độ chính xác phân lớp.

Cách tiếp cận bootstrapping để giải quyết các vấn đề này. Nó bắt đầu từ một tập nhỏ của các mẫu huấn luyến có nhãn. Bằng cách sử dụng kết hợp phương pháp huấn luyện, hai thuật toán phân lớp thống kê được sử dụng để huấn luyện và chú thích các mẫu không có nhãn, thuật toán chú thích thành công một tập dữ liệu lớn. Từ thực nghiệm chỉ ra rằng, hiệu quả tra cứu cải thiện lên 10% độ chính xác tra cứu khi được so với SVM (400 ảnh có nhãn cho huấn luyện), với các mẫu huấn luyện có nhãn ít hơn (chỉ có 20 nhãn). Bên cạnh các thuật toán được đề cập ở đây, kỹ thuật cây quyết định (decision tree) cũng được dùng để sinh các đặc trưng ngữ nghĩa.

Phương pháp cây quyết định như ID3, C4.5, CART xây dựng một cấu trúc cây bằng phân hoạch đệ quy không gian thuộc tính đầu vào thành một tập hợp không gian không chồng chéo.

1.5.3. Một số nghiên cứu theo hướng tiếp cận học không giám sát

Không giống với học có giám sát khi dữ liệu có nhãn hay có hướng dẫn trong suốt quá trình học, với học không giám sát để tra cứu ảnh dựa trên nội dung [60], dữ liệu không có nhãn, nhiệm vụ từ những đặc trưng đầu vào như vậy cần tổ chức hoặc nhóm lại.

Phân cụm ảnh là kỹ thuật điển hình của học không giám sát đối với mục đích tra cứu [43, 44, 45]. Nó dự định nhóm một bộ hình ảnh theo cách tối đa hóa độ tương tự của các đối tượng trong cụm và tối thiểu độ tương tự giữa các cụm khác nhau. Mỗi kết quả phân cụm kết hợp một nhãn và ảnh trong cùng cụm là tương tự với nhau. Thuật toán phân cụm K-means truyền thống và biến

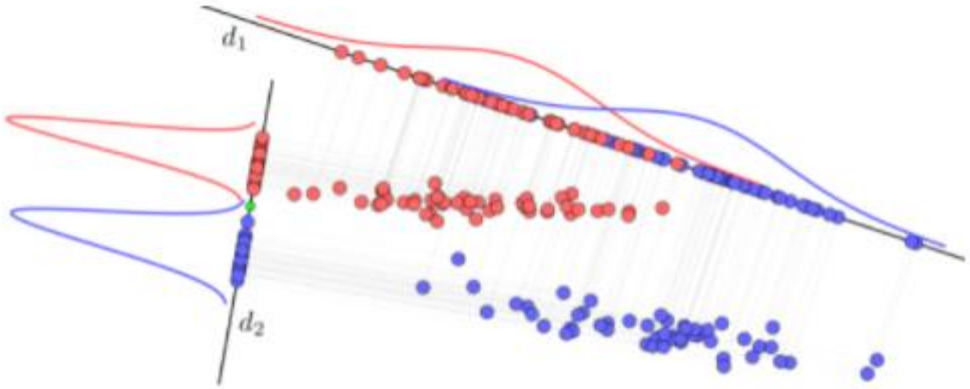
thể của nó thường được dùng để phân cụm. Trong [33, 46, 47], áp dụng thuật toán phân cụm K-means trên đặc trưng mầu mức thấp của tập ảnh huấn luyện. Sau đó, đo sự khác nhau trong mỗi cụm để sinh ra một tập chỉ mục giữa đặc trưng trực quan mức thấp và đặc tính văn bản tối ưu (từ khóa) của mỗi cụm tương ứng. Các luật chỉ mục được sinh có thể được sử dụng thêm để lập chỉ mục cho ảnh không có nhãn thêm vào ảnh cơ sở dữ liệu. Trong đề xuất phương pháp chú thích ảnh cơ sở dữ liệu tự động cho mục đích tra cứu, đầu tiên hệ thống phân cụm ảnh thành các vùng sử dụng một biến thể của K-means (PCK-means). Số lượng các cụm được thiết lập là 30. Sau đó, xác suất của mỗi khái niệm (khái niệm được định nghĩa cho cơ sở dữ liệu ảnh được sử dụng) cho một vùng được sinh ra bằng cách sử dụng Phương pháp Bayesian [48, 49]. Do đó, một hình ảnh có thể được chú thích bằng cách chọn khái niệm mà có xác suất cao nhất. Do sự phân bố phức tạp của dữ liệu ảnh (các điểm dữ liệu được lấy mẫu từ không gian đa tạp), các phương pháp truyền thống như phân cụm K-means thường không thể phân tách ảnh tốt với nhiều khái niệm khác nhau [33, 46, 47].

Để giải quyết vấn đề ở trên, phương pháp phân cụm phổ [50] đề xuất và thành công trên nhiều ứng dụng như phân đoạn ảnh, phân cụm ảnh. Một phiên bản mở rộng của N_Cut có sẵn trong [A. Y. Ng, M. I. Jordan, and Y. Weiss. 2001]. Trong phương pháp CLUE được đề xuất để giảm —khoảng cách ngữ nghĩal trong CBIR. Không giống với các hệ thống CBIR khác hiển thị danh sách các ảnh kết quả ban đầu có độ tương tự cao tới người dùng, hệ thống này cố 25 gắng lấy các cụm ảnh ngữ nghĩa. Khi đưa vào một ảnh truy vấn, một tập các ảnh mục tiêu tương tự với ảnh truy vấn được chọn là lân cận của ảnh truy vấn. Dựa vào giả thuyết rằng các ảnh có cùng ngữ nghĩa có xu hướng nằm trong cùng một cụm, phân cụm N_Cut được sử dụng để phân cụm ảnh mục tiêu thành các nhóm ngữ nghĩa khác nhau. Sau đó hệ thống sẽ hiển thị các cụm ảnh đó và điều chỉnh mô hình độ đo tương tự theo phản hồi của người dùng. Mặc dù rất

thành công trên nhiều ứng dụng như phân đoạn ảnh, phân cụm ảnh. Một phiên bản mở rộng của N_Cut có sẵn trong [A. Y. Ng, M. I. Jordan, and Y. Weiss. 2001]. Trong phương pháp CLUE được đề xuất để giảm —khoảng cách ngữ nghĩal trong CBIR. Không giống với các hệ thống CBIR khác hiển thị danh sách các ảnh kết quả ban đầu có độ tương tự cao tới người dùng, hệ thống này cố 25 gắng lấy các cụm ảnh ngữ nghĩa. Khi đưa vào một ảnh truy vấn, một tập các ảnh mục tiêu tương tự với ảnh truy vấn được chọn là lân cận của ảnh truy vấn. Dựa vào giả thuyết rằng các ảnh có cùng ngữ nghĩa có xu hướng nằm trong cùng một cụm, phân cụm N_Cut được sử dụng để phân cụm ảnh mục tiêu thành các nhóm ngữ nghĩa khác nhau. Sau đó hệ thống sẽ hiển thị các cụm ảnh đó và điều chỉnh mô hình độ đo tương tự theo phản hồi của người dùng. Mặc dù rất thành công trong phân cụm dữ liệu đa tạp, N_Cut không cung cấp một hàm chỉ mục hoàn hảo nên phương CLUE chưa đem lại kết quả tốt.

1.6. Phân tích phân biệt tuyến tính

PCA là phương pháp giảm chiều dữ liệu sao cho lượng thông tin về dữ liệu, thể hiện ở tổng phương sai, được giữ lại là nhiều nhất. Tuy nhiên, trong nhiều trường hợp, ta không cần giữ lại lượng thông tin lớn nhất mà chỉ cần giữ lại thông tin cần thiết cho riêng bài toán. Xét ví dụ về bài toán phân lớp với 2 lớp được mô tả trong Hình 3.



Hình I. 2. PCA cho bài toán phân lớp với 2 lớp

Trên Hình I.2, dữ liệu được chiếu lên các đường thẳng khác nhau. Có hai lớp dữ liệu minh hoạ bởi các điểm màu xanh và đỏ. Dữ liệu được giảm số chiều về một bằng cách chiếu chúng lên các đường thẳng khác nhau d_1 và d_2. Trong hai cách chiều này, phương của d_1 gần giống với phương của thành phần chính thứ nhất của dữ liệu, phương của d_2 gần với thành phần phụ của dữ liệu nếu dùng PCA. Khi chiếu lên d_1, các điểm màu đỏ và xanh bị chồng lấn lên nhau, khiến cho việc phân loại dữ liệu là không khả thi trên đường thẳng này. Ngược lại, khi được chiếu lên d_2, dữ liệu của hai lớp được chia thành các cụm tương ứng tách biệt nhau, khiến cho việc phân loại trở nên đơn giản hơn và hiệu quả hơn. Các đường cong hình chuông thể hiện xấp xỉ phân bố xác suất của dữ liệu hình chiếu trong mỗi lớp.

Trong Hình I.2, ta giả sử rằng dữ liệu được chiếu lên một đường thẳng và mỗi điểm được đại diện bởi hình chiếu của nó lên đường thẳng kia. Như vậy, từ dữ liệu nhiều chiều, ta đã giảm nó về một chiều. Câu hỏi đặt ra là, đường thẳng cần có phương như thế nào để hình chiếu của dữ liệu trên đường thẳng này *giúp ích cho việc phân loại tốt nhất*?. Phân loại với khoảng cách phi số liệu để truy xuất hình ảnh và biểu diễn lớp [51]. Việc Phân loại đơn giản nhất có thể được hiểu là việc tìm ra một ngưỡng giúp phân tách hai lớp một cách đơn giản và đạt kết quả tốt nhất.

Xét hai đường thẳng d_1 và d_2. Trong đó phương của d_1 gần với phương của thành phần chính nếu dùng PCA, phương của d_2 gần với phương của thành phần phụ tìm được bằng PCA. Nếu giảm chiều dữ liệu bằng PCA, sẽ thu được dữ liệu gần với các điểm được chiếu lên d_1. Lúc này việc phân tách hai lớp trở nên phức tạp vì các điểm đại diện cho hai lớp chồng lấn lên nhau. Ngược lại, nếu ta chiếu dữ liệu lên đường thẳng gần với thành phần phụ tìm được bởi PCA, tức d_2, các điểm hình chiếu nằm hoàn toàn về hai phía khác nhau của điểm màu lục trên đường thẳng này. Với bài toán phân loại, việc chiếu dữ liệu lên d_2 vì vậy

sẽ mang lại hiệu quả hơn. Việc phân loại một điểm dữ liệu mới sẽ được xác định nhanh chóng bằng cách so sánh hình chiếu của nó lên d_2 với điểm màu xanh lục này.

Qua ví dụ trên ta thấy, không phải việc giữ lại thông tin nhiều nhất sẽ luôn mang lại kết quả tốt nhất. Hai đường thẳng d_1 và d_2 trên đây không vuông góc với nhau, nghiên cứu sinh chỉ chọn ra hai hướng gần với các thành phần chính và phụ của dữ liệu để minh hoạ. Phân tích phân biệt tuyến tính (LDA -Linear Tách Analysis) ra đời nhằm giải quyết vấn đề này. LDA là một phương pháp giảm chiều dữ liệu cho bài toán phân loại. LDA có thể được coi là một phương pháp giảm chiều dữ liệu (dimensionality reduction), và cũng có thể được coi là một phương pháp phân lớp (phân lớp), và cũng có thể được áp dụng đồng thời cho cả hai, tức giảm chiều dữ liệu sao cho việc phân lớp hiệu quả nhất. Số chiều của dữ liệu mới là nhỏ hơn hoặc bằng C−1 trong đó C là số lượng lớp. Từ 'tách' được hiểu là *những thông tin đặc trưng cho mỗi lớp, khiến nó không bị lẫn với các lớp khác*. Từ 'tuyến tính' được dùng vì cách giảm chiều dữ liệu được thực hiện bởi một ma trận chiếu (projection matrix), là một phép biến đổi tuyến tính (linear transform). Trong mục dưới đây, nghiên cứu sinh sẽ trình bày về trường hợp hai lớp, tức có 2 lớp.

1.6.1. Phân tích phân biệt tuyến tính cho bài toán với hai lớp

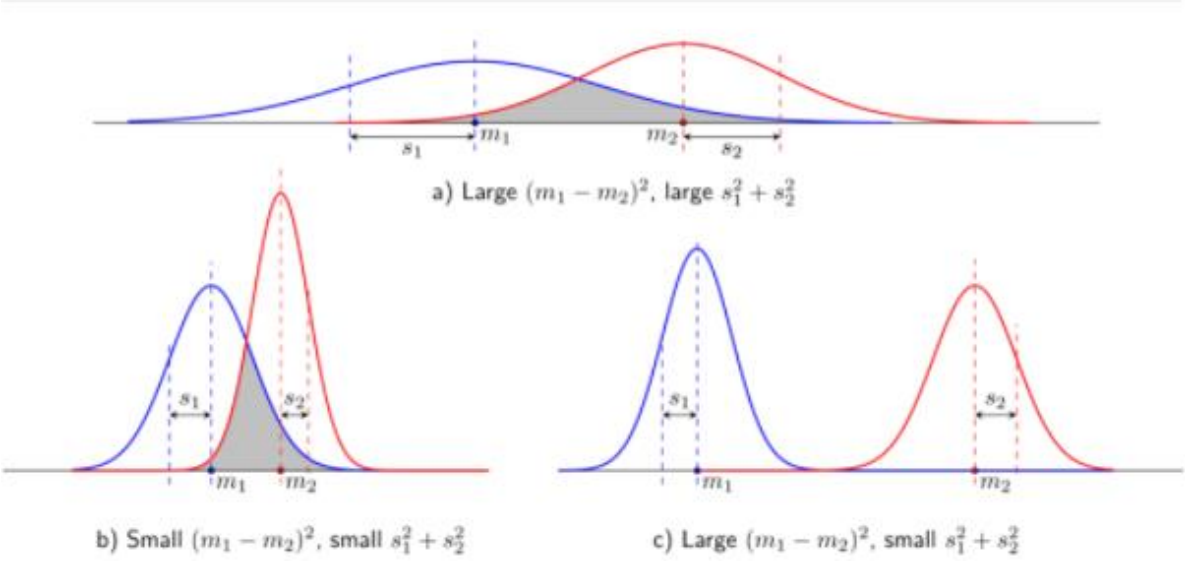
1.6.1.1 Ý tưởng cơ bản

Mọi phương pháp phân loại [52] đều được bắt đầu với bài toán nhị phân phân loại, và LDA cũng không phải ngoại lệ.

Quay lại với Hình I. 2, các đường hình chuông thể hiện đồ thị của các hàm mật độ xác suất (probability density function - pdf) của dữ liệu được chiếu xuống theo từng lớp. *Phân phối chuẩn ở đây được sử dụng như là một đại diện, dữ liệu không nhất thiết luôn phải tuân theo phân phối chuẩn.*

Độ rộng của mỗi đường hình chuông thể hiện độ lệch chuẩn của dữ liệu. Dữ liệu càng tập trung thì độ lệch chuẩn càng nhỏ, càng phân tán thì độ lệch chuẩn càng cao. Khi được chiếu lên d_1, dữ liệu của hai lớp bị phân tán quá nhiều, khiến cho chúng bị trộn lẫn vào nhau. Khi được chiếu lên d_2, mỗi lớp đều có độ lệch chuẩn nhỏ, khiến cho dữ liệu trong từng class tập trung hơn, dẫn đến kết quả tốt hơn.

Tuy nhiên, việc độ lệch chuẩn nhỏ trong mỗi lớp chưa đủ để đảm bảo độ *Tách* của dữ liệu. Xét các ví dụ trong Hình 4.



Hình I. 3. Khoảng cách phân kỳ giữa các kỳ vọng và tổng các phương sai ảnh hưởng tới độ tách của dữ liệu.

Trên Hình I.3, khoảng cách giữa các kỳ vọng và tổng các phương sai ảnh hưởng tới độ *tách* của dữ liệu. a) Khoảng cách giữa hai kỳ vọng là lớn nhưng phương sai trong mỗi lớp cũng lớn, khiến cho hai phân phối chồng lấn lên nhau (phần màu xám). b) Phương sai cho mỗi lớp là rất nhỏ nhưng hai kỳ vọng quá gần nhau, khiến khó phân biệt 2 lớp. c) Khi phương sai đủ nhỏ và khoảng cách giữa hai kỳ vọng đủ lớn, ta thấy rằng dữ liệu *tách* hơn.

Hình 3I.a) giống với dữ liệu khi chiếu lên d_1 ở Hình 1. Cả hai lớp đều quá phân tán khiến cho tỉ lệ chồng lấn (phần diện tích màu xám) là lớn, tức dữ liệu chưa thực sự *phân biệt*.

Hình I.3b) là trường hợp khi độ lệch chuẩn của hai lớp đều nhỏ, tức dữ liệu tập trung hơn. Tuy nhiên, vấn đề với trường hợp này là khoảng cách giữa hai lớp được đo bằng khoảng cách giữa hai kỳ vọng m_1 và m_2, là quá nhỏ, khiến cho phần chồng lấn cũng chiếm môt tỉ lệ lớn, và tất nhiên, cũng không tốt cho phân loại.

Hình I.3c) là trường hợp khi hai độ lệch chuẩn là nhỏ và khoảng cách giữa hai kỳ vọng là lớn, phần chống lấn nhỏ không đáng kể.

Ở đây, cần lưu ý: độ lệch chuẩn và khoảng cách giữa hai kỳ vọng đại diện cho các tiêu chí gì:

- Như đã nói, độ lệch chuẩn nhỏ thể hiện việc dữ liệu ít phân tán. Điều này có nghĩa là dữ liệu trong mỗi lớp có xu hướng giống nhau. Hai phương sai s_1^2, s_2^2 còn được gọi là các phương sai trong cùng lớp (within-class variances).
- Khoảng cách giữa các kỳ vọng là lớn chứng tỏ rằng hai lớp nằm xa nhau, tức dữ liệu giữa các lớp là khác nhau nhiều. Bình phương khoảng cách giữa hai kỳ vọng $(m_1-m_2)^2$ còn được gọi là phương sai liên lớp (between-class variance).

Hai lớp được gọi là phân biệt (*discriminative)* nếu hai lớp đó cách xa nhau (between-class variance lớn) và dữ liệu trong mỗi lớp có xu hướng giống nhau (within-class variance nhỏ). Phân tích phân biệt tuyến tính là thuật toán đi tìm một phép chiếu sao cho tỉ lệ giữa *between-clas variance* và *within-class variance* lớn nhất có thể. Phân tích phân biệt tuyến tính thưa thớt mạnh mẽ cũng được các nhà khoa học quan tâm đặc biệt.

1.6.1.2. Xây dựng hàm mục tiêu

Giả sử rằng có N điểm dữ liệu $x_1, x_2, .., x_N$ trong đó $N_1 < N$ điểm đầu tiên thuộc lớp thứ nhất, $N_2 = N - N_1$ điểm cuối cùng thuộc lớp thứ hai.

Ký hiệu $C_1 = \{n/1 \leq n \leq N_1\}$ là tập hợp các chỉ số của các điểm thuộc lớp 1 và $C_2 = \{m/N_1+1 \leq m \leq N\}$ là tập hợp các chỉ số của các điểm thuộc lớp 2. Phép chiếu dữ liệu xuống một đường thẳng có thể được mô tả bằng một véc tơ hệ số **w**, giá trị tương ứng của mỗi điểm dữ liệu mới được cho bởi:

$$y_n = w^T x_n, 1 \leq n \leq N$$

Véc tơ kỳ vọng của mỗi lớp:

$$m_k = \frac{1}{N_k} \sum_{n \in C_k} x_n, k = 1, 2 \quad (1.34)$$

Khi đó:

$$m_1 - m_2 = \frac{1}{N_1} \sum_{i \in C_1} y_i - \frac{1}{N_2} \sum_{i \in C_2} y_i = w^T (m_1 - m_2) \quad (1.35)$$

Các phương sai trong cùng lớp được định nghĩa là:

$$s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2, k = 1, 2 \quad (1.36)$$

LDA là thuật toán đi tìm giá trị lớn nhất của hàm mục tiêu:

$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 - s_2^2} \quad (1.37)$$

Tiếp theo, sẽ đi tìm biểu thức phụ thuộc giữa tử số và mẫu số trong vế phải của (4) vào **w**.

Với tử số:

$$(\mathbf{m_1} - \mathbf{m_2})^2 = \mathbf{w^T} \underbrace{(\mathbf{m_1} - \mathbf{m_2})(\mathbf{m_1} - \mathbf{m_2})^\mathbf{T}}_{\boldsymbol{S_B}} \mathbf{w} = \mathbf{w^T S_B w} \quad (1.38)$$

$\boldsymbol{S_B}$ còn được gọi là ma trận tương quan 2 lớp. Đây là một ma trận đối xứng nửa xác định dương.

Với mẫu số:

$$s_1^2 + s_2^2 = \sum_{k=1}^{2} \sum_{n \in C_k} \left(w^T (x_n - m_k)\right)^2 =$$

$$w^T \underbrace{\sum_{k=1}^{2} \sum_{n \in C_k} (x_n - m_k)(x_n - m_k)^T}_{} w = w^T S_w w \quad (1.39)$$

$\mathbf{S_W}$ còn được gọi là ma trận tương quan trong lớp. Đây cũng là một ma trận đối xứng nửa xác định dương vì nó là tổng của hai ma trận đối xứng nửa xác định dương.

Trong (5) và (6), ta đã sử dụng đẳng thức:

$$\mathbf{(a^T b)^2 = (a^T b)(a^T b) = a^T b b^T a}$$

với a, b là hai véc tơ cùng chiều bất kỳ.

Như vậy, bài toán tối ưu cho LDA trở thành:

$$\mathbf{w} = arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^T \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}} \quad (1.40)$$

1.6.2. Nghiệm của bài toán tối ưu

Nghiệm **w** của (7) sẽ là nghiệm của phương trình đạo hàm hàm mục tiêu bằng 0.

Sử dụng luật xích (chain rule) cho đạo hàm hàm nhiều biến [41] và công thức:

$\boldsymbol{\nabla_w w A_w = 2Aw}$ nếu $\boldsymbol{A}$ là một ma trận đối xứng, ta có:

$$\boldsymbol{\nabla_w J(w)} = \frac{\mathbf{1}}{(w^T S_w w)^2} (\mathbf{2 S_w w}(w^T S_w w) - \mathbf{2 w^T S_w} w^T S_w w) \quad (1.41)$$

$$\Leftrightarrow S_B w = \frac{w^T S_B w}{w^T S_w w} S_w w \quad (1.42)$$

$$= J(w) w \quad (1.43)$$

Lưu ý : Trong (1.42), ta đã giả sử rằng ma trận $\mathbf{S_W}$ là khả nghịch. Điều này không luôn luôn đúng, nhưng có một mẹo nhỏ là ta có thể xấp xỉ S_W bởi $\overline{\boldsymbol{S}}\boldsymbol{w} \approx \boldsymbol{S_w} + \boldsymbol{\lambda I}$ với λ là một số thực dương nhỏ.

Ma trận mới này là khả nghịch vì trị riêng nhỏ nhất của nó bằng với trị riêng nhỏ nhất của S_W cộng với λ tức không nhỏ hơn $\lambda > 0$. Điều này được suy ra từ việc $\mathbf{S_W}$ là một ma trận nửa xác định dương. Từ đó suy ra $\overline{\boldsymbol{S}}\boldsymbol{w}$ là một ma

trận xác định dương vì mọi trị riêng của nó là thực dương, và vì thế, nó khả nghịch. Khi tính toán, ta có thể sử dụng nghịch đảo của $\bar{S}w$.

Kỹ thuật này được sử dụng rất nhiều khi ta cần sử dụng nghịch đảo của một ma trận nửa xác định dương và chưa biết nó có thực sự là xác định dương hay không.

Quay trở lại với (1.40), vì $J(w)$ là một số vô hướng, ta suy ra **w** phải là một véc tơ riêng của $S_b S_w$ ứng với một trị riêng nào đó. Hơn nữa, giá trị của trị riêng này bằng với $J(w)$. Vậy, để hàm mục tiêu là lớn nhất thì $J(w)$ chính là trị riêng lớn nhất của $S_B^{-1} S_w$. Dấu bằng xảy ra khi **w** là véc tơ riêng ứng với trị riêng lớn nhất đó.

Từ có thể thấy ngay rằng nếu **w** là nghiệm của (1.42) thì k**w** cũng là nghiệm với *k* là số thực khác không bất kỳ. Vậy ta có thể chọn w sao cho:

$(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{w} = \mathbf{J}(\mathbf{w}) = \mathbf{L}$ = trị riêng lớn nhất của $S_w^{-1} S_w$

Khi đó, thay định nghĩa của S_B vào ta có:

$$L_w = S_w^{-1}(m_1 - m_2)\underbrace{(m_1 - m_2)^T w} = L S_w^{-1} S_w$$

Điều này có nghĩa là ta có thể chọn:

$$L_w = \alpha S_w^{-1}(m_1 - m_2) \qquad (1.44)$$

với $\alpha \neq 0$ bất kỳ.

Biểu thức (1.44) còn được biết như là Fisher's linear tách, được đặt theo tên nhà khoa học Ronald Fisher.

1.7. Thiết lập chỉ số định lượng véc tơ đối với đặc trưng

Đặc trưng lập chỉ mục đã được biết đến như một kỹ thuật quan trọng cho phép truy xuất nhanh và đối sánh các đối tượng trực quan trong thị giác máy tính. Ứng dụng tích cực nhất của đặc trưng lập chỉ mục có lẽ liên quan đến tìm kiếm lân cận gần nhất (ANN) nhanh. Trong tài liệu, các cách tiếp cận phổ biến cho vấn đề này có thể được liệt kê là phương pháp phân vùng không gian (thông

thường, KD-tree [5] và KD-tree ngẫu nhiên [26], hoặc LM-tree [24]), các phương pháp băm (chẳng hạn như LSH [8], Kerneedly LSH [12]), các phương pháp phân cụm phân cấp (chẳng hạn như từ vựng K-mean tree [19], POC-tree [22]).

Gần đây, lượng tử hóa (PQ) [9] đã được nghiên cứu tích cực cho các ứng dụng của nó trong tìm kiếm gần đúng nhanh (ANN) và lập chỉ mục đặc trưng. Các biến thể khác nhau của kỹ thuật PQ đã được trình bày để tối ưu hóa giai đoạn lượng tử hóa, chẳng hạn như PQ được tối ưu hóa [6, 20], PQ được tối ưu hóa cục bộ [10], hoặc PQ nhạy cảm với phân phối (DSPQ) [13]. PQ cũng có thể kết hợp với ý tưởng phân cụm phân cấp để tăng hiệu năng tìm kiếm như được trình bày trong [23, 25]. Các thực nghiệm mở rộng đã được tiến hành trong [23, 25], cho thấy kết quả của cây PQ và K-mean kết hợp khi so sánh với các cách tiếp cận hiện có.

Trong chương này, nghiên cứu sinh đề xuất một cách sử dụng khác của ý tưởng PQ. Đối với PQ, không gian dữ liệu đầu tiên được phân chia thành các không gian con rời rạc. Không giống như PQ, các véc tơ con của một số các không gian con liên tiếp được nhóm lại trước khi thực hiện lượng tử hóa véc tơ. Ý tưởng mới này giúp khai thác tốt hơn mối tương quan của dữ liệu trên các không gian con. Bằng cách này, một số không gian con sẽ chia sẻ một bộ định lượng chung có số lượng trọng tâm là cao hơn so với những người sử dụng trong PQ. Cụ thể, số lượng trọng tậm hoặc từ mã được sử dụng trong phương pháp của nghiên cứu sinh tỷ lệ với số lượng không gian con được nhóm lại. Mặc dù đề xuất phương pháp sử dụng số lượng từ mã cao hơn cho mỗi bộ định lượng, tổng số trọng tâm vẫn giống như trong phương pháp PQ và do đó nó tiêu thụ cùng một ngân sách bit.

Ngoài ra, luận án cũng trình bày một cách thức mới để chia không gian thành các không gian con rời rạc. Cụ thể, nghiên cứu sinh tận dụng lợi thế của

tri thức về không gian đặc trưng (tức là SIFT và GIST trong trường hợp của nghiên cứu sinh) để tìm ra không gian được tối ưu hóa. Cách thức này giúp cải thiện chất lượng mã hóa mặc dù nó không được áp dụng cho một tập dữ liệu chung.

1.8. Nghiên cứu liên quan định lượng véc tơ đối với đặc trưng và chỉ số

Vì nghiên cứu hiện tại tập trung vào các kỹ thuật lượng tử hóa, chúng ta sẽ thảo luận về các phương pháp hiện đại liên quan đến cây phân cụm và lượng tử hóa. Phân cụm theo thứ bậc. Tóm lại, cây phân cụm phân cấp được tạo ra bằng cách chia lặp lại tập dữ liệu cơ bản thành các tập con hoặc cụm nhỏ hơn bằng cách sử dụng một tiêu chuẩn thuật toán phân cụm, thường là K-mean [15], K-medoids [11], hoặc DBSCAN [4]. Từng cụm sau đó tiếp tục được chia thành các cụm con cho đến khi các tập con thu được là đủ nhỏ. Để đại diện cho sự phân rã phân cụm có thứ bậc này, một cấu trúc cây được sử dụng trong đó gốc tương ứng với tập dữ liệu gốc và mỗi nút bên trong đại diện cho một cụm con ở mức độ phân hủy tương ứng. Mỗi nút của cây cũng chứa các thông tin, chẳng hạn như, trọng tâm của các điểm dữ liệu được gán cho nút này. Đặc biệt, các nút lá của cây chứa thông tin phong phú hơn bao gồm cả các trọng tâm và các chỉ mục của các điểm dữ liệu.

Một trong những cấu trúc phân cụm phân cấp đầu tiên được nghiên cứu sinh giới thiệu trong [19], cụ thể là thuật toán K-mean tree. Việc xây dựng cây được thực hiện chính xác như mô tả trước đây trong đó thuật toán K-mean được chọn để phân cụm các các giá trị tính năng thành các nhóm nhỏ hơn.

Khi cây được tạo, nó có thể được sử dụng để thực hiện tìm kiếm xấp xỉ lân cận gần nhất (ANN) đã đưa ra một truy vấn. Tìm kiếm bắt đầu từ nút gốc và đi xuống cây, tại mỗi nút trong, nó chọn một nút con để khám phá thêm bằng cách chọn nút có chênh lệch Euclide nhỏ nhất tới truy vấn. Khi định vị tại một nút lá, quét tuyến tính được thực hiện để tìm ra ứng viên tốt nhất. Lần tìm ngược

sau đó được gọi để tinh chỉnh thêm câu trả lời tốt nhất. Tìm kiếm có thể kết thúc sớm bằng cách đặt ra hạn chế về số lượng nút tối đa được truy cập.

Đã có những tiến bộ khác nhau trong việc cải thiện cây từ vựng K-mean. Rõ ràng rằng, các tác giả trong [16, 18] đã cung cấp cho cây từ vựng K-mean một chiến lược tìm kiếm ưu tiên trong đó tìm kiếm luôn chọn nút từ đầu hàng đợi ban đầu. Hàng đợi chứa danh sách các nút ứng cử viên được lưu trữ theo thứ tự tăng dần của chênh lệch so với truy truy vấn. Các thực nghiệm trên diện rộng đã cho thấy kết quả vượt trội của cấu trúc cây này. Trong [22], các tác giả đề xuất một biến thể của cây phân cụm được gọi là Cụm được tối ưu hóa theo cặp (POC-tree). Mỗi POC-tree được tạo theo cách tương tự như từ vựng K-mean tree nhưng khác ở chỗ nó tối đa hóa không gian phân tách của mọi cặp cụm ở mỗi bước phân hủy. Do đó, sự phân hủy kết quả tương ứng với một biểu diễn của toàn bộ không gian dữ liệu. Hơn nữa, nhiều cây POC có thể được kết hợp để cải thiện hơn nữa hiệu năng tìm kiếm, đặc biệt là đối với không gian đặc trưng nhiều chiều.

Lượng tử hóa

Lượng tử hóa (PQ) [9] là một cách tiếp cận mạnh để mã hóa và biểu diễn dữ liệu. Về bản chất, PQ chia không gian dữ liệu thành các không gian con rời rạc và định lượng chúng một cách riêng biệt. Cụ thể, đối với mỗi không gian con, một bộ định lượng phụ đã học bằng cách sử dụng bộ định lượng véc tơ như K-mean. Một khi các bộ lượng tử con đã được huấn luyện, chúng được sử dụng để ánh xạ một véc tơ đầu vào thành các mã ngắn, mỗi mã tương ứng với chỉ số của một từ mã phụ của một bộ định lượng con cụ thể. Các mã ngắn sau đó được nối với từ mã hoàn chỉnh của các giá trị tính năng. Để hỗ trợ ANN, PQ đã được kết hợp với cấu trúc tệp đảo ngược để đạt được tìm kiếm [9]. Sự kết hợp có được đã cho thấy kết quả khá thú vị.

Các nghiên cứu trong [6, 10, 20], đã được trình bày với sự tập trung đặc biệt vào cải thiện chất lượng mã hóa của quá trình lượng tử hóa. Mục tiêu chính của những nghiên cứu đó là huấn luyện các bộ lượng tử để thu được một cách thích ứng sự phân bố nội tại của dữ liệu. Với mục đích này, nghiên cứu trong [20] giới thiệu hai kỹ thuật lượng tử hóa, cụ thể là ck-means và ok-means, tối ưu hóa bước huấn luyện bằng cách làm cho dữ liệu xoay vòng. Do đó, mô hình được tối ưu hóa giúp giảm đáng kể lượng tử hóa các lỗi. Tương tự, một nghiên cứu khác, được gọi là lượng tử hóa được tối ưu hóa (OPQ) [6], đã được trình bày cùng một lúc và chia sẻ ý tưởng chung. Ở đây, OPQ đầu tiên cho phép dữ liệu được căn chỉnh bằng cách sử dụng PCA và sau đó sắp xếp lại các chiều để cùng đạt được hai mục tiêu độc lập và cân bằng giữa các không gian con. Hai đặc trưng đó rất quan trọng khi huấn luyện các bộ định lượng và chúng đã được giả định là đúng với kỹ thuật PQ. OPQ hoạt động rất tốt đối với các bộ dữ liệu phân phối mô hình đơn lẻ nhưng ít hiệu quả hơn đối với trường hợp không gian đặc trưng nhiều mô hình. Trong trường hợp sau, OPQ cục bộ (LOPQ) [10] là một lựa chọn hợp lý vì nó áp dụng OPQ cục bộ trên từng cụm điểm dữ liệu. Tuy nhiên, quá trình tối ưu hóa được thực hiện mà không tính đến bước lượng tử hóa véc tơ.

Gần đây, các tác giả trong [25] đề xuất kết hợp khả năng của PQ và phân cụm thứ bậc dẫn đến một kỹ thuật mới, cụ thể là Nhúng lượng tử hóa =(EPQ). Tương tự như PQ, EQP chia không gian dữ liệu thành các không gian con rời rạc. Tuy nhiên, đối với mỗi không gian con, một cây K-mean là từ vựng riêng biệt được tạo ra để thu sự phân bố của các dữ liệu. Cây cũng có thể hoạt động như một bộ định lượng con và một cấu trúc tệp đảo ngược. Hiệu năng tìm kiếm đã được chứng minh là rất ấn tượng so với các kỹ thuật khác. Một biến thể khác của EPQ đã được giới thiệu sau đó trong [23] bởi cùng các tác giả để tiếp tục cải thiện tốc độ tìm kiếm.

Kỹ thuật lượng tử hóa độ dài đầy đủ. Điều thú vị là một số tác phẩm gần đây được trình bày trong [1, 3, 28, 29] không đưa ra giả định nào về tính độc lập lẫn nhau như được đặt ra trong kỹ thuật dựa vào PQ. Trong [1], các tác giả trình bày một kỹ thuật mới, đó là Lượng tử hóa cộng (AQ), mã hóa mỗi véc tơ đầu vào là tổng của m từ mã đến từ m sách mã.

Không giống như PQ, các từ mã có độ dài đầy đủ như các véc tơ ban đầu và do đó phân rã không gian không được thực hiện ở đây. Một phương pháp mã hóa độ dài đầy đủ khác đã được trình bày trong [3] bởi cùng các tác giả mã hóa từng véc tơ bằng phương pháp cây lượng tử hóa (TQ). Mỗi nút của cây tương ứng với một bản ghi mã, trong khi một cạnh được liên kết với một chiều. Kết quả là, mỗi cuốn sách mã quản lý lượng tử hóa cho một vài chiều đã được chỉ định cho nó. Các chiều còn lại được đặt thành 0, dẫn đến một véc tơ lượng tử hóa khá thưa. Một phiên bản tối ưu hóa của kỹ thuật này (OTQ) cũng đã được trình bày, nơi dữ liệu được xoay vòng để đạt được mô hình phù hợp hơn. Các thực nghiệm đã báo cáo rằng, hai kỹ thuật vượt trội so với PQ về chất lượng mã hóa nhưng ít hiệu quả về thời gian hơn IVFADC [9].

Để giảm chi phí tính toán của các kỹ thuật AQ, TQ và OTQ, nghiên cứu trong [28] áp đặt một ràng buộc bổ sung cho mô hình lượng tử hóa. Đặc biệt, nó giả định rằng tích của các phần tử liên từ điển là một giá trị không đổi. Giả định bổ sung này làm cho quá trình tính toán chênh lệch giữa một truy vấn và các từ mã trở nên đơn giản. Hơn nữa, các tác giả cũng giới thiệu một biến thể của kỹ thuật này khai thác tính chất thưa của các từ mã khi huấn luyện các bộ mã [29]. Mô hình thu được đã được hiển thị rất hiệu quả khi so sánh với các phương pháp khác.

1.9. Cách tiếp cận được đề xuất định lượng véc tơ đối với đặc trưng và chỉ số

1.9.1. Lượng tử hóa véc tơ và lượng tử hóa véc tơ con

Trong tài liệu [6, 7, 9], lượng tử hóa véc tơ (VQ) đề cập đến một quá trình xây dựng một cuốn sách mã C bao gồm K từ mã $\{c_1, c_2, \ldots, c_K\}$, và ánh xạ một điểm dữ liệu đầu vào đã cho $x \in R^D$ tới từ mã gần nhất. Về mặt hình thức, ánh xạ được ký hiệu là $q(x)$ như sau:

$$q(x) \leftarrow \underset{c_k \in C}{\operatorname{argmin}}\, d(x, c_k) \tag{1.56}$$

trong đó $\mathbf{d}(x, c_k)$ biểu thị chênh lệch Euclide và $q(x)$ được gọi là bộ lượng tử. Để khẳng định chất lượng của bộ lượng tử hóa, lỗi lượng tử hóa thường được sử dụng để đo độ méo của việc lượng tử hóa một tập dữ liệu X bao gồm n điểm dữ liệu trong không gian R^D

$$E = \frac{1}{n}\sum_{x \in X}\|x - q(x)\|^2 \tag{1.57}$$

ở đây || q - p || biểu thị chuẩn Euclide giữa hai điểm p và q. Theo quy ước, [9] thường sử dụng ký hiệu $\|.\|_2$ để tính toán các lỗi lượng tử hóa hoặc cấu trúc. Do đó, nghiên cứu sinh đã sử dụng ký hiệu này trong luận án của mình.

Sổ mã C có thể thu được bằng cách áp dụng thuật toán phân cụm (ví dụ: K-mean) phân chia tập hợp các điểm dữ liệu trong không gian R^D thành K cụm, mỗi cụm được biểu diễn bằng một trọng tâm hoặc từ mã. Bằng cách lượng tử hóa véc tơ, một véc tơ đầu vào có thể được biểu diễn bằng một đoạn mã rất ngắn với ngân sách bit gần bằng log(K). Do đó, VQ đã được sử dụng rộng rãi để xử lý nhiều ứng dụng thị giác máy tính như túi từ trực quan (Bag-of-visual-words) [27], tìm kiếm lân cận gần nhất nhanh [18].

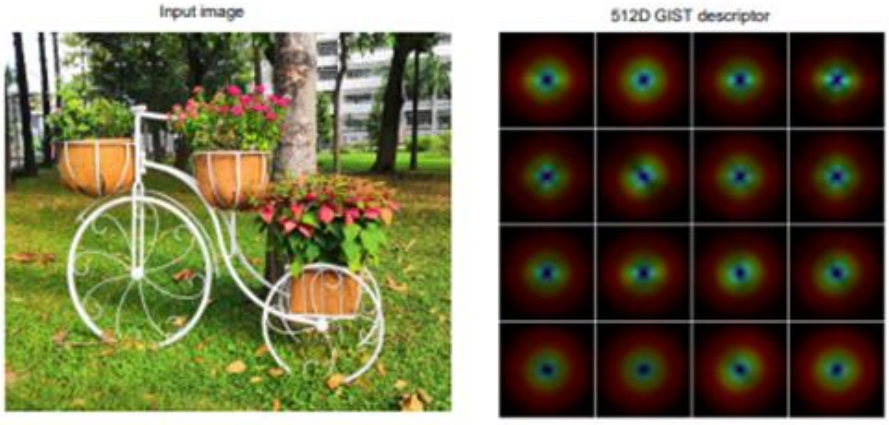
Mặt khác, lượng tử hóa (PQ) [9] mở rộng ý tưởng về VQ trong đó không gian dữ liệu ban đầu được chia thành m không gian con riêng biệt, tiếp theo là kỹ thuật VQ được áp dụng riêng cho các véc tơ con của mỗi không gian con.

Gọi $a_i(x)$ là véc tơ thứ J^{th} của véc tơ đầu vào $x \in R^D$ và C_j là tập mã con được xây dựng từ không gian con thứ J^{th} (ở đây j = *1,2,..,m*), một bộ lượng tử con $q_j(y)$ ánh xạ một véc tơ đầu vào $x \in R^{D/m}$ tới từ mã con gần nhất của C_j bởi.

$$q_j\ (y) \leftarrow arg\ \min_{c_{j,k} \in C_j} \mathbf{d}(x, c_{j,k}) \qquad (1.58)$$

trong đó $c_{j,k}$ là từ mã con thứ k của C_j

Trong cách tiếp cận PQ, mỗi sổ mã con thường bao gồm *K* từ mã con. Giá trị cao hơn của K, là yếu tố quyết định sự phân hủy của dữ liệu. Kết quả là, độ méo lượng tử hóa thấp hơn với chi phí tăng ngân sách bit. Hơn nữa, vì VQ được áp dụng riêng biệt cho từng không gian con, nên mối tương quan của dữ liệu giữa các không gian con không được khai thác, dẫn đến biểu diễn dư thừa của các trọng tâm con. Ví dụ, chúng ta hãy xem xét GIST 512D, nhiều véc tơ con (tức là các khối vuông của bộ mô tả trong Hình 1) trong hai không gian con liên tiếp khá giống nhau. Các véc tơ con đó tương ứng với mô tả trực quan về các khối nền liền kề của một hình ảnh. Bằng cách xây dựng các sổ mã con riêng biệt, các véc tơ con đó thuộc về các trọng tâm con của các không gian con khác nhau. Do đó, nhiều trọng tâm con tương tự được tạo ra.



Hình I. 4. Hình ảnh đầu vào (bên trái) và bộ mô tả GIST 512D của nó (bên phải). Nhiều phần nền trong hình ảnh giống nhau về nội dung trực quan dẫn đến sự giống nhau của các khối mô tả.

Để giải quyết vấn đề này, nghiên cứu sinh đề xuất tập hợp các véc tơ con của các không gian con liên tục thành các nhóm lớn hơn trước khi áp dụng lượng tử hóa véc tơ. Cụ thể, phương pháp được đề xuất, gọi là lượng tử hóa véc tơ con (PSVQ), kết hợp dữ liệu từ h, với $1 < h \leq m$, không gian con liên tục và thực hiện lượng tử hóa véc tơ để tạo $m^* = m/h$ con- máy lượng tử. Mỗi cái bao gồm $K^* = h \times K$ trọng tâm con. Làm như vậy, một số không gian con sẽ chia sẻ cùng một bộ định lượng con và do đó tạo ra sự phân rã mịn hơn của dữ liệu cơ bản. Cần lưu ý rằng PSVQ không làm tăng tổng số từ mã phụ (tức là tổng số vẫn có $m^* \times K^* = m \times K$ từ mã phụ). Lượng tử hóa một véc tơ con $y = a_j(x)$ thuộc không gian con thứ j hiện được xử lý bởi $q_i^*(y)$), với $i = [j/h] + 1$ và do đó $1 \leq i \leq m^*$, như sau:

$$q_i^*(y) \leftarrow arg \min_{c_i, k \in C_i^*} d(y, c_i, k) \tag{1.59}$$

trong đó $1 \leq k \leq K^*$, và C_i^* là sổ mã con được huấn luyện trên tập dữ liệu bao gồm $n \times h$ điểm dữ liệu, thu được bằng cách nhóm các véc tơ con từ không gian con thứ (s + 1) thành không gian con thứ (s + h) th với $s = (i - 1) \times h$.

Toàn bộ quá trình được mô tả ở trên có thể được xây dựng trên Thuật toán 1. Thuật toán nhận đầu vào là tập dữ liệu X và một số tham số, sau đó thực hiện quá trình huấn luyện để tạo $m *$ sách mã con C_i^*, mỗi trong số đó chứa K^* từ mã con. Về bản chất, khía cạnh tính toán chính của Thuật toán 1 dựa trên thuật toán K-mean. Do đó, độ phức tạp về thời gian và bộ nhớ của nó tương đương với K-means. Cần lưu ý rằng vòng lặp chính trong Thuật toán 1 bị giới hạn bởi $m * = m / h$, là một giá trị nhỏ ($m = 8$ trong tất cả các thử nghiệm của nghiên cứu sinh). Ngoài ra, vì quá trình huấn luyện được thực hiện ngoại tuyến, do đó, chi phí về thời gian không phải là mối quan tâm lớn.

$$E^* = \frac{1}{n} \sum_{x \in X} \cdot \sum_{j=1}^{m} \left\| a_j(x) - q_i^*\left(a_j(x)\right) \right\|^2 \tag{1.60}$$

trong đó $i = [j/h] + 1$ như được mô tả trước đây

Thuật toán 1 Huấn luyện PSVQ *(m, X, h, K*)*

1: **Input**: *m số lượng sổ mã con, X là tập dữ liệu để huấn luyện các danh sách mã con* PSVQ, *h* số chênh lệch con được hợp nhất và K* số lượng trọng tâm con trong mỗi sổ mã.

2: **Output**: Danh sách các mã sách con *m/h*.

3: $m^* \leftarrow m/h$

4: $L \leftarrow \emptyset$

5: Tách X thành m tập con: $X_1, X_2, \ldots, Xm$

6: **for** $i = 1$ **to** m^* **do**

7: $s \leftarrow (\mathrm{i} - 1) \times \mathrm{h}$

8: $Y \leftarrow$ nhập h tập con liên tục $\{X_{s+1}, X_{s+2}, \ldots, X_{s+h}\}$

9: $C_i^* \leftarrow$ Kmeans(Y, K*) {Áp dụng K-means trên tập con Y để tạo K^* trọng tâm con. Ở đây, C_i^* chứa K^*trọng tâm con kết quả}

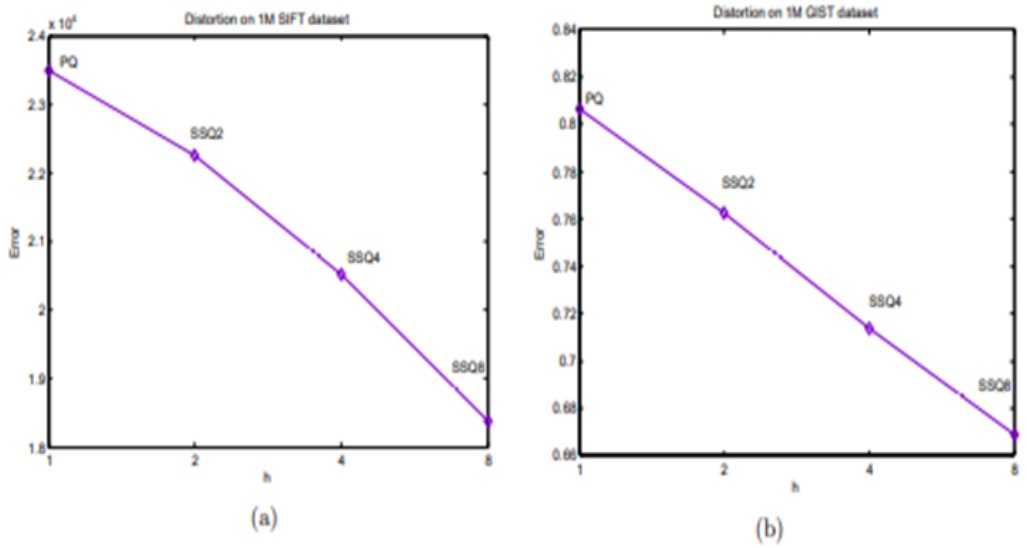
10: $L \leftarrow$ Append(L, C_i^*) {nối thêm C_i^* vào L}

11: **end for**

12: return L {L là một mảng của các sách mã con m^*, mỗi trong chúng chứa K^* trọng tâm con}

Với sự phân rã không gian mới này, dữ liệu được trình bày ở mức tốt hơn vì các véc tơ con tương tự trong các không gian con khác nhau có khả năng được biểu diễn bằng một trọng tâm chung. Do đó, nghiên cứu sinh khai thác nhiều hơn mối tương quan của dữ liệu trên các không gian con và làm cho phù hợp hơn với dữ liệu, điều này cuối cùng dẫn đến độ méo mã hóa thấp hơn. Cụ thể, chúng ta sẽ chỉ ra rằng sai số lượng tử hóa (E*) tỷ lệ nghịch với giá trị của h. Với mục đích này, nghiên cứu sinh đã huấn luyện các bộ lượng tử con trên bộ dữ liệu huấn luyện (cho SIFT và GIST [9]) và sau đó tính toán các lỗi cho các bộ cơ sở dữ liệu (n = 1000000 điểm dữ liệu). Ba giá trị khác nhau của h (tức là số lượng không gian con cho lượng tử hóa hay gọi tắt là SSQ) được xem

xét trong thực nghiệm này, bao gồm h = 2 (SSQ2), h = 4 (SSQ4), h = 8 (SSQ8) và cuối cùng nghiên cứu sinh đặt m = 8 là một lựa chọn phổ biến trong tài liệu [9, 20]. Rõ ràng, trường hợp h = 1 tương ứng với kỹ thuật PQ chuẩn. Như có thể thấy trong Hình 10, lỗi lượng tử hóa giảm đáng kể khi chúng ta chuyển từ PQ (tức là h = 1) sang SSQ2, SSQ4 và SSQ8 cho cả tập dữ liệu SIFT và GIST.



Hình I. 5. Lỗi lượng tử hoá cho tập dữ liệu 1M SIFT(a) và 1M GIST (b).

Sau khi huấn luyện các bảng ghi mã, chúng có thể được sử dụng để lượng tử hóa một cơ sở dữ liệu đặc trưng nhất định. Gọi G là cơ sở dữ liệu gồm n điểm trong không gian R^d và G_q là mã lượng tử hóa của G. Mỗi mã lượng tử hóa của G_q thu được bằng cách lượng tử hóa một điểm dữ liệu $x \in G$ dựa trên các sổ mã con C_i^*. Cụ thể hơn, trước tiên chúng ta chia x thành m véc tơ con $a_j(x)$ với $1 \leq j \leq m$, và sau đó tìm một từ mã con của C_i^* $(i = [j/h] + 1)$ là gần nhất với $a_j(x)$ về chênh lệch Euclide. Chỉ số liên quan đến từ mã con thu được, được coi là lượng tử hóa con của $a_j(x)$. Kết quả là, mã lượng tử hóa hoàn chỉnh của x là một véc tơ nguyên có kích thước m mà mỗi phần tử có giá trị trong khoảng $0, 1, \ldots, K * -1$. Lặp lại quá trình trên với mọi điểm dữ liệu trong Q,

chúng ta thu được cơ sở dữ liệu lượng tử hóa G^q có kích thước $n \times m$ và tiêu tốn ít dung lượng bộ nhớ hơn nhiều so với cơ sở dữ liệu gốc G. Việc tìm kiếm bây giờ có thể được thực hiện bằng cách chọn một chênh lệch thích hợp. Vì mục đích này, tính toán chênh lệch không đối xứng (ADC) thường được sử dụng trong tài liệu [9, 20]. Theo chênh lệch ADC, có nghĩa là chúng ta xấp xỉ chênh lệch giữa hai điểm x và r trong không gian R^d bằng chênh lệch giữa x và $q^*(r)$ trong đó $q^*(r)$ là một ánh xạ lượng tử hóa nối các lượng tử hóa con. mã như sau:

$$q^*(r) \leftarrow \left\{q^*\left(a_j(r)\right)\right\} \tag{1.61}$$

Trong đó $j = 1, 2, \ldots, m$ và $i = [j/h] + 1$

Với việc sử dụng chênh lệch ADC, quá trình tìm kiếm được nêu trên Thuật toán 2. Thuật toán nhận đầu vào là véc tơ truy vấn $x \in R^d$, tham số R cho biết số lượng câu trả lời ứng viên và cơ sở dữ liệu lượng tử hóa G_q. Về cơ bản, tìm kiếm được thực hiện bởi một tìm kiếm toàn diện, quét mọi phần tử của Gq, tính toán chênh lệch ADC tương ứng và cập nhật danh sách ngắn chứa R câu trả lời hay nhất được tìm thấy cho đến nay. Quá trình duy trì danh sách ngắn có thể được thực hiện một cách hiệu quả bằng cách sử dụng cấu trúc Maxheap [9].

Mục tiêu chính của Thuật toán 2 là thể hiện chất lượng mã hóa của phương pháp lượng tử hóa PSVQ của nghiên cứu sinh. Với mục đích này, khía cạnh thời gian chạy không được giải quyết một cách tối ưu ở đây và thuật toán hoạt động như một tìm kiếm thô. Do đó, độ phức tạp tính toán là tuyến tính với kích thước của cơ sở dữ liệu G_q. Cần lưu ý rằng bước tính toán chuyên sâu nhất của Thuật toán 2 liên quan đến tính toán chênh lệch ADC. Tuy nhiên, độ phức tạp tính toán của bước này bị giới hạn bởi số lượng từ mã con *(i.e., $m \times K^*$)* là một giá trị tương đối nhỏ. Do đó, nghiên cứu sinh có thể tính toán trước chênh lệch giữa x và tất cả các từ mã con, lưu trữ chúng vào bảng tra cứu 1D và truy

cập bảng để tính toán hiệu quả chênh lệch ADC. Kết quả của Thuật toán 2 được so sánh với các phương pháp mã hóa khác như PQ và ck-mean. Hiệu năng chi tiết sẽ được trình bày trong các thực nghiệm tiếp theo của nghiên cứu sinh.

Algorithm 2 ADCSearch*(x, R, G_q)*

1: **Input:** một truy vấn đầu vào (x), số các câu trả lời ứng viên được trả về *(R),* và cơ sở dữ liệu được lượng hóa (G_q).
2: **Output:** Một danh sách ngắn của R phần tử gần nhất với x sử dụng chênh lệch ADC.
3: $L_R \leftarrow \emptyset$
4: $n \leftarrow$ length of (G_q)
5: **for** $i :=$ 1 **to** n **do**
6: $d \leftarrow \mathbf{d}(x, c_i)$ {Tính chênh lệch ADC giữa x và mã lượng hóa a, ở đây c_i biểu thị các từ mã con kết hợp với mã lượng hóa thứ i của Gq}
7: $L_R \leftarrow$ MaxHeap*(R, d, i)* {cập nhật danh sách chứa *R* câu trả lời tốt nhất}
8: **end for**
9: return L_R

1.9.2. Phân vùng không gian

Khi thực hiện phân rã không gian trong kỹ thuật PQ, giả định rằng các không gian con là độc lập lẫn nhau. Tuy nhiên, điều này không phải lúc nào cũng đúng với bất kỳ tập dữ liệu nào ngoại trừ SIFT [14]. Trong bộ mô tả SIFT, các giá trị tính năng được tính toán cho từng vùng cục bộ trong đó vùng được chia thành các cửa sổ con 4 × 4, mỗi cửa sổ trong số đó là một biểu đồ 8-bin về các hướng được tính cho các pixel được gán cho nó. Các biểu đồ sau đó được đóng gói để tạo thành các các giá trị tính năng 128D. Do đó, nếu chúng ta chia các véc tơ thành m không gian con (tức là m = 8, 16), các không gian con thu được hầu như độc lập do thứ tự tự nhiên của các thứ nguyên SIFT. Tương tự, GIST cũng được tính toán với nguyên tắc tương tự như minh họa trong [21]. Đặc biệt, tương ứng với tập dữ liệu GIST1M1, GIST được tính toán cho một

hình ảnh đầu vào nhất định, như sau: thay đổi kích thước hình ảnh theo kích thước cố định, áp dụng lưới 4 × 4 cho mỗi trong ba kênh màu, tính toán biểu đồ định hướng Gaborbased 16-bin cho mỗi ô trong số 16 ô lưới, và cuối cùng nối các biểu đồ kết quả để tạo thành GIST 960D (tức là 3 × 16 × 20). Tuy nhiên, tập dữ liệu GIST1M của 960D đã được tổ chức lại với một chút sửa đổi theo thứ tự nối các biểu đồ. Do đó, thứ tự tự nhiên của các kích thước trong tập dữ liệu GIST1M không phản ánh chính xác cách xây dựng GIST.

Hình I. 6. Hình ảnh đầu vào (bên trái) và bộ mô tả SIFT được tính toán tại 4 điểm chính (bên phải)

Xem xét các khía cạnh đã nói ở trên, nghiên cứu sinh lập luận rằng với một số kiến thức trước đây về không gian đối tượng, chúng ta có thể sử dụng một chiến lược phân rã không gian phù hợp. Quan sát này cũng đã được khai thác trong kỹ thuật PQ, nơi các tác giả đề xuất sử dụng "thứ tự có cấu trúc" để chia nhỏ các tập dữ liệu. Theo thứ tự có cấu trúc, điều đó có nghĩa là tất cả các kích thước có cùng modulo chỉ số m (m = 8 trong công trình nghiên cứu của nghiên cứu sinh) được nhóm vào một không gian con. Kết quả thử nghiệm cho thấy rằng ý tưởng này mang lại sự cải thiện đáng kể về chất lượng mã hóa cho GIST.

Trong nghiên cứu hiện tại, nghiên cứu sinh đề xuất phân chia không gian đối tượng theo cách giống như khi chúng ta nối biểu đồ của các đối tượng SIFT

và GIST. Như đã mô tả trước đây, đối với cả đặc trưng SIFT và GIST, vùng hình ảnh đầu vào được chia thành các vùng con 4 × 4 và sau đó nội dung hình ảnh thống kê được thu thập riêng cho từng vùng con. Kết quả là, có rất ít mối tương quan giữa biểu đồ của hai vùng con. Tóm lại, nghiên cứu sinh chia không gian đối tượng thành m không gian con (m = 8 trong thử nghiệm của nghiên cứu sinh) trong đó mỗi không gian con chứa các kích thước đã được gán cho biểu đồ được tính từ hai vùng con liên tục cho cả đối tượng SIFT và GIST. Tất nhiên, nếu chúng ta đặt m = 16, sự phân rã không gian sẽ thích hợp vì mỗi không gian con tương ứng với một vùng con của hình ảnh đầu vào. Tuy nhiên, nghiên cứu sinh đã chọn m = 8 để so sánh với các kỹ thuật PQ tiên tiến nhất.

1.10. Thí nghiệm định lượng véc tơ đối với đặc trưng và chỉ số

1.10.1. Bộ dữ liệu và cài đặt

Để xác nhận hiệu năng của phương pháp được đề xuất, nghiên cứu sinh đã tiến hành hai thực nghiệm khác nhau. Thực nghiệm đầu tiên nhằm mục đích so sánh chất lượng mã hóa hoặc hiệu suất lượng tử hóa của phương pháp được đề xuất với các kỹ thuật dựa trên PQ khác có mối quan tâm đặc biệt đến việc tối ưu hóa bộ lượng tử. Với mục đích này, nghiên cứu sinh đã chọn hai phương pháp đại diện bao gồm PQ tiêu chuẩn và phương pháp ck-mean [20].

Thử nghiệm thứ hai so sánh hiệu quả tìm kiếm hoặc hiệu năng tìm kiếm ANN của phương pháp đề xuất. Vì mục tiêu chính ở đây là kiểm tra khía cạnh lập chỉ mục, do đó nghiên cứu sinh đã chọn các chiến lược lập chỉ mục tốt nhất bao gồm: IVFADC (tức là, PQ kết hợp với cấu trúc tệp đảo ngược sử dụng chênh lệch ADC [9]), cây K-mean [16] (tức là, kỹ thuật hiệu quả nhất để tìm kiếm ANN của thư viện FLANN2), POC-tree [22], EPQ [25] và EPQ được tối ưu hóa (OEPQ) [23]. Các mã nguồn nằm trong môi trường C/C ++ và quá trình kiểm tra được thực hiện trên máy tính tiêu chuẩn có cấu hình sau: Windows 10, RAM 16Gb, Intel Core (Dual-Core) i7 2.1 GHz.

Đối với các chỉ số đánh giá, thử nghiệm đầu tiên sử dụng Recall @ R, trong khi thử nghiệm thứ hai sử dụng các đường cong tốc độ / độ chính xác. Những tiêu chí này thường được sử dụng trong tài liệu cho các nhiệm vụ tương tự [9, 17, 18, 20]. Cụ thể, Recall @ R đo lường nhóm các câu trả lời đã sửa từ danh sách các phần tử R rút gọn. Việc tăng tốc độ được tính toán tương đối thông qua tìm kiếm quét tuyến tính để tránh tác động của cấu hình máy tính. Để có một báo cáo ổn định, nghiên cứu sinh đã chạy hàng nghìn truy vấn và sau đó tính kết quả trung bình làm đầu ra cuối cùng.

Hai tập dữ liệu công khai được sử dụng cho tất cả các thử nghiệm của nghiên cứu sinh, bao gồm ANN SIFT1M và ANN GIST1M [9]. Một số thông tin cơ bản được báo cáo trong Bảng 3.1 cho cả hai bộ dữ liệu. Để huấn luyện các bộ lượng tử, PQ, CK-means và phương pháp được đề xuất sử dụng bộ huấn luyện khác với bộ cơ sở dữ liệu và bộ kiểm tra. Ngoài ra, tất cả các phương thức này cũng sử dụng các cài đặt tham số giống nhau như sau: m = 8 và K = 256. Khác biệt, các phương thức còn lại (ví dụ: K-mean tree, POC-tree, EPQ và OEPQ) sử dụng trực tiếp cơ sở dữ liệu để huấn luyện các bộ định lượng (nếu có) và cả cây phân cụm vì chúng đã được nhắm mục tiêu cụ thể đến khía cạnh lập chỉ mục.

Dataset	#Training set	#Database	#Queries	#Dimension
ANN_SIFT1M	100,000	1,000,000	10,000	128
ANN_GIST1M	500,000	1,000,000	1000	960

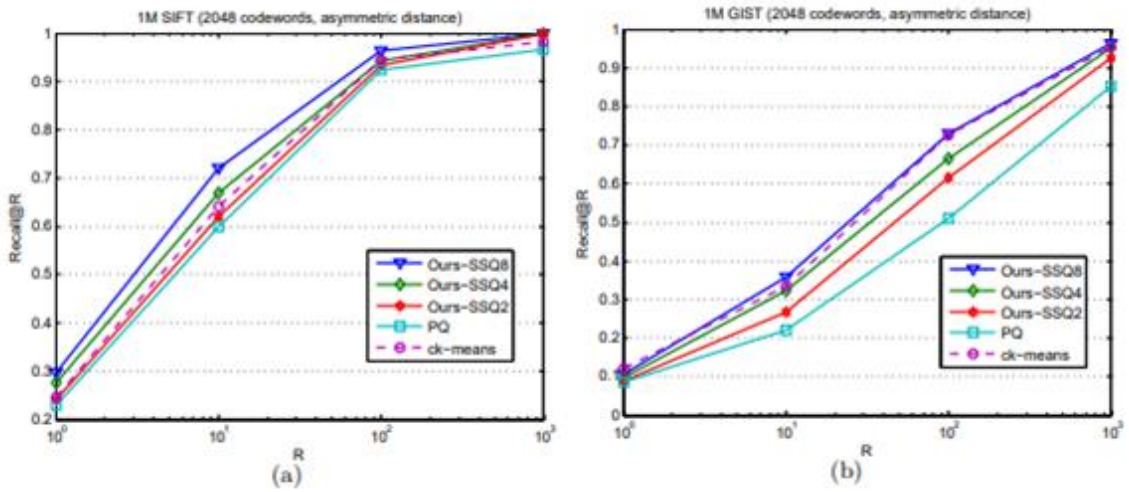
Bảng I. 1. Các bộ lọc dữ liệu được sử dụng trong các thí nghiệm của NCS

1.10.2. Đánh giá chất lượng mã hóa

Nghiên cứu sinh so sánh phương pháp được đề xuất với PQ và ck-mean về chất lượng mã hóa cho các hoàn cảnh khác nhau của h như đã đề cập trước đây, bao gồm SSQ2, SSQ4 và SSQ8 tương ứng với 4, 2 và 1 bộ lượng tử con,

tương ứng. Đối với các trường hợp, sử dụng h = 2 (SSQ2) có nghĩa là chúng ta nhóm dữ liệu của hai không gian con liên tiếp trước khi áp dụng quá trình lượng tử hóa véc tơ và do đó tạo ra tổng cộng 4 lượng tử con. Với mục đích này, Thuật toán 2 được sử dụng để tính toán điểm số Recall @ R của thuật toán của nghiên cứu sinh. Hình 4 trình bày kết quả của tất cả các phương pháp cho cả tập dữ liệu SIFT và GIST. Chúng ta có thể quan sát thấy rằng SSQ2, SSQ4 và SSQ8 hoạt động tốt hơn cả PQ và ck-means cho SIFT. Đối với tập dữ liệu GIST, ck-mean chỉ vượt trội hơn so với biến thể SSQ8 của nghiên cứu sinh có thể do tác động của chiều cao. Những kết quả này khá ấn tượng khi xem xét rằng ck-means đã được được tối ưu hóa để nắm bắt phân phối nội tại của dữ liệu cơ bản.

Mặt khác, Hình 12 cũng cho thấy rằng PQ khi kết hợp với chiến lược thứ tự có cấu trúc để phân rã không gian vẫn không vượt trội so với mô hình đơn giản nhất của chúng ta là SSQ2. Rõ ràng, biến thể SSQ8 hoạt động tốt nhất vì sự tương quan của dữ liệu được khai thác tối đa trên tất cả các không gian con, dẫn đến độ méo mã hóa thấp nhất như được trình bày theo kinh nghiệm trong Hình 10. Tuy nhiên, chất lượng mã hóa vượt trội này không ngụ ý cải thiện tìm kiếm hiệu quả nhờ chi phí tính toán khi tính toán chênh lệch. Do đó, người ta phải cân bằng giữa cả hai khía cạnh này tùy thuộc vào ngữ cảnh ứng dụng cụ thể.



Hình I. 7. Chất lượng mã hoá cho SIFT (a) và GIST (b)

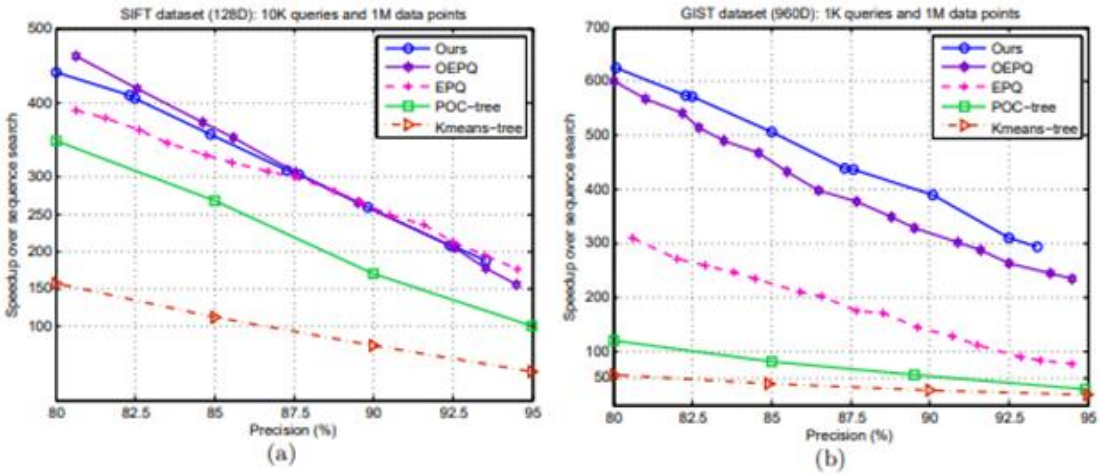
1.10.3. Đánh giá tìm kiếm xấp xỉ lân cận gần nhất

Trong phần này, nghiên cứu sinh kiểm tra hiệu quả của các bộ lượng tử con được đề xuất cho nhiệm vụ tìm kiếm ANN. Với mục đích này, nghiên cứu sinh sử dụng chiến lược lập chỉ mục như các tác giả đã trình bày trong [23]. Cách thức này là một đường dẫn tiêu chuẩn bao gồm một cây phân cụm để biểu diễn phân cấp của toàn bộ cơ sở dữ liệu và một công cụ kinh nghiệm (heuristic) để thúc đẩy tìm kiếm cho một truy vấn nhất định. Cấu trúc lập chỉ mục này được kết hợp với các bộ định lượng dựa trên PQ và đã sử dụng chênh lệch ADC để trả về danh sách các câu trả lời ứng viên có thứ tự. Theo hiểu biết của nghiên cứu sinh, phương pháp này hiện mang lại kết quả hiện đại nhất cho tìm kiếm ANN trên các tập dữ liệu đã nghiên cứu. Tuy nhiên, những kết quả này thu được với một hạn chế là không gian dữ liệu cho các đặc trưng GIST được chia thành nhiều không gian con nhỏ để tránh vấn đề ảnh hưởng nhiều chiều.

Khác với công việc đã đề cập ở trên, mục tiêu chính của nghiên cứu này là đánh giá hiệu quả tìm kiếm khi sử dụng các cài đặt thậm chí rất đơn giản (tức là không gian dữ liệu được chia thành 8 không gian con cho cả hai đặc trưng SIFT ad GIST). Như đã đề cập trước đây, hiệu suất tìm kiếm ANN được đo

bằng các đường cong tăng tốc/ độ chính xác. Ở đây, độ chính xác tìm kiếm có thể được coi là một trường hợp đặc biệt của Recall @ R trong đó R = 1. Điều đó có nghĩa là phần nhỏ các câu trả lời chính xác khi xem xét danh sách ngắn chỉ chứa một ứng cử viên. Để có được một loạt các điểm tăng tốc độ chính xác, một số tham số sẽ khác nhau khi thực hiện truy vấn. Chi tiết hơn được đề cập đến nguyên tác [23].

Hình 5 so sánh hiệu quả tìm kiếm ANN của phương pháp của nghiên cứu sinh với các nghiên cứu gần đây bao gồm EPQ, OEPQ, POC-tree và Kmeans-tree. Như thể hiện trong Hình 5, phương pháp được đề xuất hoạt động ngang bằng với các kỹ thuật lập chỉ mục hiện đại (tức là OEPQ và EPQ) cho SIFT và đặc biệt là vượt trội hơn tất cả các phương pháp khác cho tập dữ liệu GIST. Trung bình, phương pháp được đề xuất nhanh hơn khoảng 2,5 × đến 3,0 × so với thuật toán tốt nhất của thư viện FLANN (tức là Kmeans-tree). Mức tăng tốc thậm chí còn cao hơn đối với GIST. Thu một ảnh nhanh cụ thể với độ chính xác tìm kiếm là 85%, thuật toán của nghiên cứu sinh nhanh hơn 500 × và 365 × so với tìm kiếm theo trình tự cho GIST và SIFT, tương ứng. Rõ ràng, chúng ta thậm chí có thể thu được các kết quả thú vị hơn nữa bằng cách chia không gian dữ liệu thành hơn 8 không gian con (m = 16 chẳng hạn). Tuy nhiên, nghiên cứu sinh đã không tiến hành tùy chọn này trong công việc hiện tại.



Hình I. 8. Hiệu suất tìm kiếm ANN cho SIFT (a) và GIST (b)

Cuối cùng, nghiên cứu sinh cũng cung cấp một đánh giá ngắn về phương pháp của nghiên cứu sinh với kỹ thuật IVFADC [9] bằng cách so sánh tốc độ tìm kiếm tương đối với thuật toán tốt nhất của FLANN như được đề xuất trong [23]. Một ưu điểm của cách này là có thể tránh được rào cản và đánh giá chủ quan do thực hiện lại thuật toán và điều chỉnh tham số. Theo báo cáo của các tác giả trong [9] và [23], tốc độ của IVFADC cao hơn khoảng 1,5 × đến 2,0 × so với thuật toán tốt nhất của FLANN đối với cùng các lựa chọn trên tập dữ liệu SIFT, trong khi phương pháp được đề xuất, như đã nhận thấy trước đây, là hiệu quả hơn khoảng 2,5 × đến 3,0 ×. Các kết quả thu được là khá ấn tượng, chứng tỏ hiệu quả của các bộ định lượng được đề xuất mặc dù ý tưởng là đơn giản về mặt khái niệm.

1.10.2. Kết luận định lượng véc tơ đối với đặc trưng và chỉ mục

Phương pháp được đề xuất đã được thiết kế để khai thác mối tương quan của dữ liệu bằng cách cho phép một trọng tâm được liên kết với nhiều hơn một không gian con. Do đó, mật độ của mỗi cụm cao hơn khi so sánh với bộ lượng tử riêng lẻ được tạo riêng biệt trong mỗi không gian con.

Ngoài ra, phương pháp đề xuất cũng xem xét kiến thức trước về phân phối dữ liệu để quyết định hành động chia nhỏ không gian phù hợp. Các bộ lượng tử kết quả có các tốt ở chỗ chúng mang lại lỗi mã hóa thấp hơn, tạo biểu diễn dữ liệu tốt hơn và sử dụng cùng một ngân sách bit để lưu trữ các từ mã con như các phương pháp PQ chuẩn. Các thực nghiệm khác nhau đã được tiến hành cho thấy hiệu quả vượt trội của phương pháp được đề xuất so với các kỹ thuật khác. Để mở rộng nghiên cứu này, nghiên cứu sinh dự định nghiên cứu việc kết hợp phương pháp học sâu để tăng hiệu năng tìm kiếm.

1.11. Kết luận chương 1

Với dữ liệu ảnh lớn như hiện nay và lượng ảnh tăng lên theo từng giờ, từng ngày, việc nghiên cứu các phương pháp CBIR hiệu quả cực kỳ cần thiết. Và đối với hệ thống CBIR việc tăng độ chính xác tra cứu ảnh và tăng tốc độ tra cứ ảnh là hai việc cần làm đầu tiên và cần thiết. Để làm được hai việc này thì hệ thống CBIR phải tập trung vào hai giai đoạn quan trọng nhất là trích rút đặc trưng và tính độ tương tự.

Trong chương này, nghiên cứu sinh đã hệ thống lại những lý thuyết cơ bản và một số nghiên cứu liên quan đến tra cứu hình ảnh dựa trên nội dung, bao gồm mô tả nội dung trực quan, đo độ tương tự/khoảng cách, lược đồ lập chỉ mục, tương tác người dùng và đánh giá hiệu năng hệ thống. các đặc trưng mức thấp gồm màu sắc, hình dạng, kết cấu và thông tin không gian; lựa chọn và trích rút đặc trưng; các kỹ thuật học máy, học sâu cho tra cứu ảnh; các độ đo tương tự cho tra cứu ảnh; tổ chức thực nghiệm và đánh giá hiệu năng. Đặc biệt, Chương này đã phân tích nghiên cứu liên quan đến các giai đoạn trong CBIR để thấy được ưu điểm và hạn chế của các nghiên cứu hiện nay. Trên cơ sở các phân tích này, định hướng nghiên cứu của luận án sẽ tập trung vào nghiên cứu các phương pháp giảm chiều thường được sử dụng là PCA, biến

đổi Karhunen – Loeve(KL), phản hồi liên quan (Relevance feedback – RF) và các phương pháp mạng nơ ron, kỹ thuật học máy, và học sâu nhằm giải quyết vấn đề thu hẹp khoảng trống ngữ nghĩa giữa các đặc trưng mức thấp và các khái niệm ngữ nghĩa mức cao để cải thiện tốc độ và độ chính xác tra cứu ảnh.

Nội dung cụ thể cần nghiên cứu ở các chương tiếp theo bao gồm:

(1) Cải thiện chất lượng của biểu diễn đặc trưng được trích rút;

(2) Khai thác các tập mẫu có liên quan và không liên quan;

(3) Giảm chiều của biểu diễn đặc trưng;

(4) Chuyển biểu diễn đặc trưng sang dạng hiệu quả;

(5) Chọn độ đo tương tự hiệu quả.

Chương 2: NÂNG CAO HIỆU QUẢ TRA CỨU ẢNH DỰA TRÊN NỘI DUNG BẰNG CÁCH KẾT HỢP KHOẢNG CÁCH TỐI ƯU VÀ PHÂN TÍCH PHÂN BIỆT TUYẾN TÍNH

Việc tra cứu ảnh dựa trên nội dung được thực hiện bằng cách so sánh sự tương tự giữa biểu diễn ảnh truy vấn và từng biểu diễn ảnh trong cơ sở dữ liệu. Do đó, biểu diễn ảnh và độ đo tương tự là hai phần cốt lõi của tra cứu ảnh dựa trên nội dung. Trong tra cứu ảnh với phản hồi liên quan, tính toán khoảng cách và phân lớp có một ảnh hưởng lớn lên độ chính xác tra cứu ảnh. Trong chương này, luận án trình bày phương pháp tra cứu ảnh đề xuất, gọi là ODLDA (Image Retrieval using the optimal distance and linear discriminant analysis). Phương pháp đề xuất có thể khai thác phản hồi của người dùng từ tập các ảnh liên quan và không liên quan, mà sử dụng phân tích phân biệt tuyến tính để tìm một chiếu tuyến tính với một độ đo tương tự cải tiến. Các kết quả thực nghiệm thực hiện trên hai tập dữ liệu tiêu chuẩn đã thấy sự tiến bộ của phương pháp đề xuất. Phương pháp đề xuất có thể khai thác hiệu quả phản hồi của người dùng từ tập hợp ảnh không liên quan, sử dụng phân tích phân biệt tuyến tính để tìm một phép chiếu tuyến tính với một số đo tương tự được cải thiện.

2.1. Giới thiệu

Trong một hệ thống CBIR tiêu biểu, các đặc trưng trực quan mức thấp bao gồm màu, kết cấu, và hình dạng, mà được trích rút tự động và được biểu diễn thành các véc tơ đặc trưng. Cũng cần lưu ý rằng, các véc tơ đặc trưng là tốt nếu chúng thể hiện ngữ nghĩa cao của ảnh và phục vụ tốt cho việc so sánh. Để tìm các ảnh mong muốn, người dùng đưa ra một ảnh mẫu và hệ thống trả lại các ảnh tương tự dựa trên các đặc trưng được trích rút. Khi hệ thống hiển thị một danh sách các ảnh mà tương tự với ảnh truy vấn, người dùng đánh dấu các ảnh liên quan nhất đến ảnh truy vấn để nhận về một danh sách phản hồi. Hệ thống

dựa vào danh sách phản hồi này để học biểu diễn hoặc độ đo tương tự của ảnh để cái thiện độ chính xác của tra cứu ảnh.

Do vậy, véc tơ đặc trưng biểu diễn ảnh và độ đo tương tự là hai nhân tố chính mà ảnh hưởng đến hiệu quả của hệ thống tra cứu ảnh. Cải thiện độ chính xác của hệ thống CBIR là một vấn đề thách thức trong nghiên cứu. Để cái tiến độ chính xác, cần giảm khoảng trống ngữ nghĩa trong hệ thống CBIR. Khoảng trống ngữ nghĩa hàm ý sự khác nhau giữa đặc trưng mức thấp và khái niệm ngữ nghĩa mức cao của ảnh. Để giảm khoảng trống này, cần áp dụng học máy vào quá trình tra cứu ảnh.

Gần đây, có các kết quả tốt do ứng dụng mạng nơ ron tích chập (CNN - convolutional neural network) vào hệ thống CBIR. Nó được chỉ ra rằng nếu một mạng CNN được huấn luyện trong một ngữ cảnh có giám sát đầy đủ trên một tập các nhiệm vụ nhận dạng đối tượng rộng, các đặc trưng được trích rút từ CNN có thể giải quyết nhiều nhiệm vụ như phân lớp đối tượng ảnh, nhận dạng cảnh, phát hiện thuộc tính, và tra cứu [4, 21, 22, 23, 53, 54, 55, 56].

Nghiên cứu trong [22] đã chỉ ra rằng hiệu năng của các hệ thống CBIR sử dụng CNN là cạnh tranh. Để cải tiến độ chính xác tra cứu ảnh, phương pháp đề xuất sử dụng mạng CNN để xây dựng một tập đặc trưng có ngữ nghĩa cao. Bên cạnh đó, phương pháp đề xuất học độ đo tương tự để có một độ đo cải tiến phù hợp với dữ liệu hơn.

Ý tưởng của học độ đo tương tự là để tìm một độ đo khoảng cách tối ưu mà cực tiểu khoảng cách giữa các cặp ảnh tương tự và cực đại khoảng cách giữa các cặp ảnh không tương tự. Sau đó, độ đo khoảng cách tối ưu này được sử dụng để phân hạng lại toàn bộ tập ảnh và trả lại các kết quả tốt hơn. Trong luận án, nghiên cứu sinh đề xuất một kỹ thuật tra cứu ảnh hiệu quả (ODLDA). Phương pháp đề xuất chính xác hơn một số phương pháp đã có bởi vì biểu diễn đặc trưng là có ngữ nghĩa cao hơn và các độ đo tương tự được học là phù hợp

với dữ liệu hơn. Bằng thực nghiệm với hai cơ sở dữ liệu tiêu chuẩn, độ chính xác của phương pháp được đề xuất được chỉ ra.

2.2. Nghiên cứu liên quan

Học độ đo tương tự trong tra cứu ảnh dựa vào nội dung đã nhận được sự chú ý của cộng đồng nghiên cứu [14, 19]. Trong tra cứu ảnh với phản hồi liên quan, dữ liệu đầu vào của các thuật toán học độ đo tương tự thường được chia thành hai nhóm: nhóm thứ nhất gồm các cặp ảnh tương tự; và nhóm thứ hai gồm các cặp ảnh tương tự và các cặp ảnh không tương tự.

Ý tưởng của việc điều chỉnh các trọng số của hàm khoảng cách đã được đưa vào trong tra cứu ảnh dựa vào nội dung như SRIR [57]. Phương pháp này tận dụng ưu điểm của thông tin từ các cặp ảnh tương tự và xem xét sự phân tán dữ liệu trên mỗi chiều để xây dựng một hàm khoảng cách euclide cải tiến.

Trong phương pháp MCML [50,58], một độ đo khoảng cách Mahalanobis sao cho các mẫu cùng lớp sẽ được ánh xạ vào cùng điểm. Bài toán học độ đo khoảng cách được xem như bài toán tối ưu lồi và được giải bởi phương pháp giảm gradient. Tuy nhiên, hạn chế của phương pháp là độ phức tạp tính toán lớn bởi vì nó sử dụng phương pháp giảm gradient để giải bài toán tối ưu lồi.

Ý tưởng của phương pháp LMNN [59] là cực tiểu khoảng cách của các mẫu cùng nhãn trong K lân cận gần nhất và cực đại khoảng cách của các mẫu mà không cùng nhãn bởi lề cực đại. Nó sử dụng hàm khoảng cách Mahalanobis. Ý tưởng này được biểu diễn như một bài toán tối ưu và được giải bởi phương pháp SDP [30] để tìm độ đo khoảng cách cải tiến.

Thuật toán trực tuyến cho học độ tương tự ảnh có thể mở rộng OASIS (Online Algorithm for Scalable Image Similarity) [32, 60] được thiết kế chuyên biệt để làm việc với các ràng buộc cặp. Tuy nhiên, chúng được dựa vào độ bền vững với dữ liệu đầu vào hoặc cấu trúc của các ràng buộc (yêu cầu dữ liệu đầu vào là các véc tơ thưa). Do đó, nó khó để áp dụng trong thực tế.

Ý tưởng trong phương pháp của Xing [14] là quy về bài toán tối ưu lồi mà cực tiểu tổng khoảng cách của các cặp ảnh tương tự với ràng buộc tổng khoảng cách của các cặp ảnh mà không tương tự là cực đại. Trong pha khởi tạo, phương pháp sử dụng hàm khoảng cách Euclide cải tiến với A=I. Phương pháp của Xing trình bày một hàm khoảng cách cải tiến trong đó A là kết quả của bài toán tối ưu lồi. Tuy nhiên, phương pháp của Xing có độ phức tạp tính toán lớn do sử dụng phương pháp gradient và chưa tận dụng thông tin của các cặp ảnh tương tự.

Ý tưởng của phương pháp RCA [32] là chỉ sử dụng các cặp ảnh tương tự, tìm một biến đổi dữ liệu dựa vào một ma trận phương sai mà được sinh ra từ các cặp ảnh tương tự. Từ đó, nó cải tiến hàm khoảng cách Mahalanobis bằng việc thay đổi ma trận trọng số. Mặc dù phương pháp này có chi phí tính toán thấp hơn của Xing, tuy nhiên, phương pháp RCA bị giới hạn là chỉ xem xét tập các ảnh tương tự.

Từ phân tích giới hạn của các nghiên cứu liên quan ở trên, luận án đề xuất một phương pháp tra cứu ảnh cải tiến. Cải tiến hàm khoảng cách dựa trên cực đại tỉ số giữa tổng khoảng cách của các cặp ảnh không tương tự và tổng khoảng cách của các cặp ảnh tương tự. Ở đây, NCS xét cả tập các ảnh tương tự và không tương tự để tìm ma trận trọng số và cái tiến độ chính xác của tra cứu ảnh.

Trong dự án truy vấn theo nội dung hình ảnh (QBIC), Nghiên cứu sinh đang nghiên cứu các phương pháp để truy vấn cơ sở dữ liệu hình ảnh trực tuyến lớn bằng cách sử dụng nội dung của hình ảnh làm cơ sở của các truy vấn. Ví dụ về nội dung Nghiên cứu sinh sử dụng bao gồm màu sắc, kết cấu và hình dạng của các đối tượng và vùng hình ảnh. Các ứng dụng tiềm năng bao gồm y tế (`Hãy cho tôi những hình ảnh khác chứa khối u có kết cấu giống như hình này'), báo ảnh (`Hãy cho tôi những hình ảnh có màu xanh lam ở trên cùng và màu đỏ ở dưới cùng'), và nhiều ứng dụng khác trong nghệ thuật, thời trang, lập danh mục,

bán lẻ và công nghiệp. Các vấn đề chính bao gồm dẫn xuất và tính toán các thuộc tính của hình ảnh và đối tượng cung cấp chức năng truy vấn hữu ích, phương pháp truy xuất dựa trên sự tương đồng chứ không phải đối sánh chính xác, truy vấn bằng ví dụ hình ảnh hoặc hình ảnh do người dùng vẽ, giao diện người dùng, tinh chỉnh truy vấn và điều hướng, cơ sở dữ liệu chiều cao lập chỉ mục, và dân số cơ sở dữ liệu tự động và bán tự động. Nghiên cứu sinh hiện có một hệ thống nguyên mẫu được viết bằng X/Motif và C chạy trên RS/6000 cho phép nhiều truy vấn khác nhau và cơ sở dữ liệu thử nghiệm gồm hơn 1000 hình ảnh và 1000 đối tượng được điền từ các hình ảnh nghệ thuật clip ảnh có sẵn trên thị trường. Trong bài báo này, chúng tôi trình bày các thuật toán chính cho kết cấu màu, hình dạng và truy vấn phác thảo mà Nghiên cứu sinh sử dụng, hiển thị các kết quả truy vấn ví dụ và thảo luận về các hướng trong tương lai.

Câu nói nổi tiếng này của nhà văn Pháp Antoine de Saint Exup´ery được áp dụng đến cuộc sống cũng như thị giác máy tính. Nhận thức của con người về hình ảnh rất nhiều vượt quá bề mặt trực quan của pixel, màu sắc và đối tượng. Ý nghĩa của một hình ảnh không thể được mô tả đơn giản bằng cách liệt kê tất cả các đối tượng chứa trong đó và xác định bố cục không gian của chúng. Chúng ta là con người có thể nắm bắt được rất nhiều sự đa dạng và thông tin phức tạp có trong một hình ảnh ngay từ cái nhìn đầu tiên, chẳng hạn như các sự kiện xảy ra trong cảnh được miêu tả, các hoạt động được thực hiện bởi những người, các mối quan hệ giữa chúng, bầu không khí và tâm trạng của hình ảnh, và cảm xúc được truyền tải bởi nó. Nhiều khái niệm trong số này không được mô tả bằng văn bản và được minh họa rõ nhất bằng cách cung cấp một hình ảnh ví dụ.

Ví dụ trong hình một sự đa dạng của thông tin được truyền tải bởi hình ảnh. Hình ảnh được mô tả ở đó có thể được mô tả từ nhiều góc độ: nội dung ngữ nghĩa, phong cách nghệ thuật của nó, những cảm xúc mà nó gợi lên trong người quan sát…

CÁC ĐỐI TƯỢNG
- Người giúp việc ≺ Phụ nữ ≺ Người
- Chiếc đầm màu đen.

Tủ quần áo ≺ Nội thất ≺ Cửa sổ
- Lâu đài Liselund ≺ Lâu đài

BỐI CẢNH
- Phòng kiểu cũ
- Phòng ngập nắng
- Người phụ nữ trước cửa sổ bên cạnh tủ quần áo
- Phòng ≺ Trong nhà

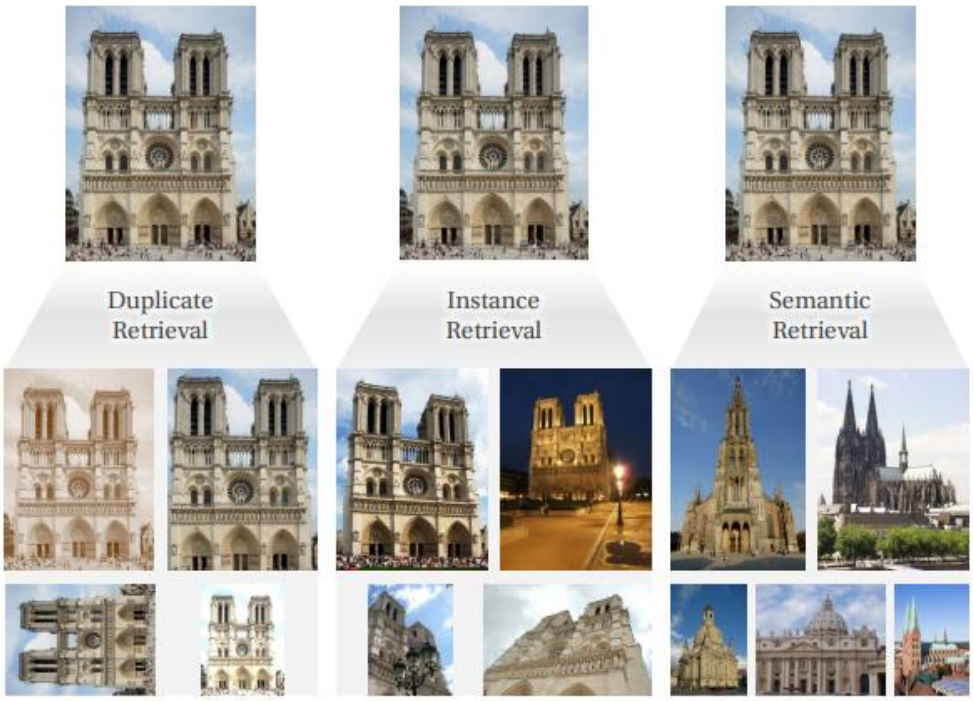
Hình II. 1 Một ví dụ về sự mơ hồ và giàu ngữ nghĩa.

Tất cả các khái niệm được liệt kê ở phía bên tay phải có thể được sử dụng để mô tả hình ảnh ở bên trái, trong khi những người quan sát khác nhau sẽ chú ý đến các tập hợp con khác nhau của các khía cạnh này. Hơn nữa, một số các khái niệm có thể được tổ chức theo thứ bậc, được biểu thị bằng dấu "≺", biểu thị mối quan hệ hyponomy ("is-a"). Siêu thông tin về chính hình ảnh đó. Tùy thuộc vào nền tảng của họ và bối cảnh tình huống, những người quan sát khác nhau sẽ cảm nhận và giải thích hình ảnh này khác nhau. Tìm kiếm hình ảnh trên web bằng mô tả văn bản hoặc từ khóa do đó chắc chắn sẽ thất bại, bởi vì hầu hết các hình ảnh đều không đầy đủ được mô tả trong văn bản xung quanh chúng, chủ yếu vì hai lý do: Thứ nhất, nó thường khó khăn, nếu không muốn nói là không thể liệt kê tất cả các khía cạnh của một hình ảnh một cách rõ ràng,

do với số lượng khả năng diễn giải có thể là vô hạn. Thứ hai, nó không phải là cần thiết để làm như vậy, vì hầu hết các khía cạnh của hình ảnh đều có sẵn trực tiếp cho

người xem chỉ bằng cách nhìn vào nó. Do đó, văn miêu tả tập trung nhất thường trên siêu thông tin không được mã hóa trong chính hình ảnh, chẳng hạn như tác giả. Ví dụ, hình ảnh hiển thị trong Hình II.1. có thể được mô tả là một bản tái tạo bằng nhiếp ảnh của bức tranh "Cửa sổ giấc mơ ngày xưa Lâu đài Liselund" của Georg Achen. Điều này sẽ ngăn không cho hình ảnh này bị được tìm thấy bởi những người dùng đang tìm kiếm hình ảnh của một người phụ nữ đang nhìn ra ngoài cửa sổ, hình ảnh thể hiện hoạt động "mơ mộng", hoặc những hình ảnh mang không khí u uất. Tìm kiếm thông qua cơ sở dữ liệu hình ảnh lớn không phải bằng từ khóa văn bản nhưng việc sử dụng một ví dụ đại diện làm truy vấn do đó là cách tự nhiên nhất, trực tiếp nhất, và cách diễn đạt để tìm hình ảnh với một nội dung cụ thể, có thể là phức tạp và khó xác định. Cách tiếp cận này được gọi là hình ảnh dựa trên nội dung hồi (CBIR) [49] và là một lĩnh vực nghiên cứu tích cực từ năm 1992 [31,36].Smeulders et al. [49] trong cuộc khảo sát mở rộng của họ vào cuối "những năm đầu" của CBIR năm 2000.

CBIR và khoảng cách ngữ nghĩa trong kỷ nguyên học sâu 3



Hình II. 2. Ví dụ về ba bộ ảnh khác nhau được truy xuất với cùng một truy vấn tuỳ thuộc vào loại nhiệm vụ CBIR

Trong hai thập kỷ đã trôi qua kể từ đó, lĩnh vực dựa trên nội dung truy xuất hình ảnh đã trải qua ít nhất hai cuộc cách mạng lớn (thêm về điều đó trong Mục 2). Tuy nhiên, hầu hết các thách thức và phương hướng chính đã được xác định hồi đó. Một trong những thách thức này là lỗ hổng ngữ nghĩa, như Smeulder et al. gọi nó đi: "Khoảng trống ngữ nghĩa là sự thiếu trùng khớp giữa các thông tin mà người ta có thể trích xuất từ dữ liệu trực quan và diễn giải rằng cùng một dữ liệu có cho một người dùng trong một tình huống nhất định." [49, giây. 2.4] Được diễn đạt bằng những lời của de Saint Exup´ery, khoảng cách ngữ nghĩa là sự khác biệt giữa việc cảm nhận một hình ảnh bằng mắt - một cách khách quan, như một sự miêu tả các vật thể, hình dạng, kết cấu - và cảm nhận hình ảnh bằng trái tim—một cách chủ quan, bao gồm kiến thức thế giới và cảm xúc, đọc "giữa các điểm ảnh".

Kích thước của khoảng cách ngữ nghĩa phụ thuộc vào mức độ trừu tượng của tìm kiếm mục tiêu mà người dùng theo đuổi. Smeulder et al. [49] xác định mức độ trừu tượng này trên phạm vi liên tục giữa hai cực của miền hẹp và miền rộng. Cái này thuật ngữ được giải thích tốt nhất trên cơ sở ba thuật ngữ phù hợp nhất hiện nay Các tác vụ CBIR, được mô tả trong Hình II. 2: Tìm kiếm truy xuất trùng lặp các hình ảnh có cùng nội dung. Những cái này là các biến thể có nguồn gốc từ cùng một bức ảnh nhưng có thể đã được hậu kỳ, được xử lý khác nhau liên quan đến cắt xén, chia tỷ lệ, điều chỉnh màu sắc, độ sáng, độ tương phản vv …Truy xuất phiên bản tìm kiếm các hình ảnh mô tả cùng một phiên bản của một đối tượng, tức là một người hoặc một tòa nhà nhất định. Nhờ bản chất của nó như là một xác định rõ ràng nhưng nhiệm vụ rất lớn với sự thật rõ ràng, đây là nhiệm vụ lớn nhất đã nghiên cứu nhiệm vụ con CBIR [4,3, 9, 20, 25, 29, 30, 38, 42,46, 48, 50]. Một số bộ dữ liệu đã được thiết lập có sẵn cho nhiệm vụ này [28,39,40,43] và tiến bộ đáng kể đã được thực hiện trong vài năm qua, mà Nghiên cứu sinh sẽ phác thảo trong Phần 2.

Truy xuất ngữ nghĩa bao gồm hầu hết phổ còn lại rộng hơn so với truy xuất cá thể và nhằm mục đích tìm kiếm các hình ảnh thuộc cùng một danh mục như truy vấn. Điều quan trọng cần lưu ý là danh mục không nhất thiết có nghĩa là lớp đối tượng trong bối cảnh này. Trong thực tế, tập hợp các danh mục có thể bị hạn chế không có gì ngoài trí tưởng tượng của người dùng và một hình ảnh duy nhất thường thuộc về với số lượng danh mục cao đáng kể cùng một lúc (xem Hình II. 1). Như vậy, các mục tiêu tìm kiếm chính xác của người dùng hiếm khi có thể được xác định dựa trên riêng hình ảnh truy vấn và gần như chắc chắn cũng sẽ khác nhau giữa những người dùng, thậm chí cho cùng một truy vấn. Vì vậy, các cách tiếp cận vấn đề này thường bao gồm tương tác với người dùng để điều chỉnh biện pháp tương tự được sử dụng bởi hệ thống đến điều đó trong tâm trí người dùng [5,7, 12,15,55].

Do đó, việc học các biểu diễn hình ảnh có ý nghĩa để nắm bắt được sự khác biệt về ngữ nghĩa tốt và các khía cạnh khác nhau của ý nghĩa của một hình ảnh là điều tối quan trọng tầm quan trọng. Mặc dù có liên quan thực tế, nhiệm vụ phụ CBIR này đã nhận được ít được chú ý hơn đáng kể so với truy xuất cá thể, chủ yếu là do ít hơn khái niệm được xác định rõ ràng về "sự liên quan" và "sự tương đồng" và kết quả là thiếu của một điểm chuẩn phù hợp. Trong công việc này, chúng tôi sẽ xem xét các phương pháp tiếp cận gần đây để truy xuất hình ảnh ngữ nghĩa và đánh giá trạng thái hiện tại của khoảng cách ngữ nghĩa, hai mươi năm sau khi kết thúc "những năm đầu" của CBIR

Truy xuất trùng lặp đánh dấu một đầu của quang phổ, vì đó là miền hẹp nhất khả thi. Trong trường hợp này, khoảng cách ngữ nghĩa hầu như không tồn tại và tất cả những gì cần khắc phục nó là một danh sách các bất biến liên quan đến nội dung của hình ảnh (ví dụ: luân canh, cắt xén…). Miền càng rộng, khoảng cách ngữ nghĩa càng lớn. Mặc dù khó khăn hơn so với truy xuất trùng lặp, nhưng truy xuất cá thể có thể vẫn được xử lý bằng cách kết hợp các mẫu hình ảnh đặc biệt chi tiết và bố cục hình học. Truy xuất hình ảnh dựa trên nội dung đã đạt được tiến bộ đáng kể trong lĩnh vực này trong hai thập kỷ qua. Tuy nhiên, khả năng áp dụng các kỹ thuật như vậy bị hạn chế đối với các khái niệm chung hơn nhiều phạm vi rộng của truy xuất ngữ nghĩa, như chúng ta thấy trong Phần 3. Một cách để khắc phục khoảng cách ngữ nghĩa này, theo Smeulders et al. [49], nằm trong việc tích hợp các nguồn thông tin ngữ nghĩa từ bên ngoài ảnh. Trong Phần 4, chúng tôi xem xét các tiếp cận theo hướng này, tiếp theo là một cuộc thảo luận về những gì vẫn còn thiếu cho thúc đẩy CBIR trong phạm vi rộng hơn nữa (Phần 5). Sự phát triển của truy xuất sơ khai Từ năm 2000 đến 2020, CBIR - với trọng tâm cụ thể là truy xuất phiên bản - đã trải qua hai sự thay đổi mô hình chính: Lần đầu tiên bắt đầu vào năm 2003 [48] và bắt đầu bằng việc điều chỉnh và cải tiến tiếp theo các kỹ thuật từ văn bản truy xuất. Làn

sóng thành tựu đột phá thứ hai bắt nguồn từ áp dụng các phương pháp học sâu cho CBIR, bắt đầu từ năm 2014 [4,45]. chúng tôi phác thảo các mốc quan trọng của hai thời kỳ đổi mới này trong những điều sau đây. Các tính năng thủ công và từ trực quan. Các tính năng địa phương như các từ trực quan Năm 2003, Sivic và Zisserman [48] đã tìm kiếm để tìm các lần xuất hiện của một đối tượng nhất định trong video và, với mục đích này, đã điều chỉnh bộ mô tả tài liệu bag-of-words (BoW), phổ biến trong lĩnh vực văn bản truy xuất, để truy xuất hình ảnh. Tương tự như các từ, họ sử dụng hình ảnh cục bộ các tính năng tại các điểm chính đặc biệt và định lượng chúng thành một từ vựng "hình ảnh từ" bằng cách sử dụng thuật toán phân cụm k-Means. Tương tự như truy xuất văn bản, số lần xuất hiện của các từ trực quan trên mỗi hình ảnh được tính và số lượng được tổng hợp thành một vectơ tf-idf đại diện cho toàn bộ hình ảnh. Vì khoảng cách Euclide là không có ý nghĩa trong không gian nhiều chiều, độ tương tự cosin sau đó được sử dụng để đánh giá sự giống nhau của hai đại diện hình ảnh như vậy. Quá trình này minh họa khuôn khổ chung để trích xuất các biểu diễn hình ảnh đã được sử dụng trong CBIR từ thời điểm đó cho đến ngày nay [30]: Một địa phương trình trích xuất tính năng tính toán các tính năng tại các điểm chính trong một hình ảnh nhất định. những địa phương này các tính năng sau đó được nhúng vào một không gian khác, chẳng hạn như các chỉ số được lượng tử hóa của từ trực quan. Cuối cùng, chúng được tổng hợp thành một đại diện toàn cầu.

Biểu diễn toàn cục cho phép truy xuất hiệu quả danh sách ban đầu của hình ảnh ứng cử viên. Ngoài ra, các tính năng cục bộ thường được sử dụng để thực hiện một bước xác minh không gian và xếp hạng lại cho các ứng cử viên xếp hạng cao nhất để loại bỏ khớp sai [48,39]. Kỹ thuật này khá cụ thể đối với truy xuất cá thể và đối sánh các vectơ đặc trưng cục bộ giữa truy vấn và hình ảnh đã truy xuất để xác minh rằng các tính năng cục bộ có bố cục hình học phù hợp.

Hướng tới các nhúng phức tạp hơn Các công việc tiếp theo của thời đại này tập trung chủ yếu vào việc cải thiện bước nhúng và tổng hợp, trong khi sử dụng cùng một trình trích xuất tính năng cục bộ trong suốt một thập kỷ. Hessian-affine trình phát hiện [34] thường được sử dụng để tìm các điểm chính mà tại đó các tính năng cục bộ sẽ được chiết xuất. Máy dò này tìm thấy điểm quan tâm không thay đổi đối với affine biến đổi cũng như mạnh mẽ để hạn chế những thay đổi của ánh sáng và quan điểm.

Sau đó, các điểm chính này được mô tả bằng các tính năng SIFT [33] hoặc RootSIFT [1]. Cái sau là một phép biến đổi đơn giản của SIFT, bao gồm L1 - bình thường hóa vectơ SIFT và lấy căn bậc hai của phần tử. Trong không gian kết quả, khoảng cách Euclide giữa các vectơ RootSIFT tương ứng với biểu đồ hạt nhân phù hợp trong không gian SIFT. Trong trường hợp của Sivic và Zisserman [48], việc nhúng biến đổi từng vectơ đặc trưng vào một không gian gồm các vectơ chỉ mục từ vựng one-hot với trọng số tf-idf. Tập hợp sau đó chỉ đơn giản bao gồm trong một hoạt động tổng hợp. Tuy nhiên, đại diện cho các vectơ đặc trưng bởi một số nguyên duy nhất (chỉ số cụm), gây ra sự mất mát nghiêm trọng về thông tin và không nắm bắt được phân phối thực tế của các tính năng Tốt. Việc gán cứng cho một cụm đơn lẻ hơn nữa không mạnh đối với các nhóm nhỏ các biến thể của bộ mô tả cục bộ gần với ranh giới cụm. Để khắc phục những vấn đề này, Perronnin et al. [38] đề xuất việc sử dụng các vectơ Fisher cho CBIR. Đào tạo dữ liệu được lượng tử hóa thành các từ trực quan bằng cách khớp mô hình hỗn hợp Gaussian. Mỗi vectơ đặc trưng cục bộ sau đó được chuyển đổi thành độ dốc của khả năng đăng nhập của nó với tôn trọng các phương tiện của Gaussian. Điều này nhận ra một nhiệm vụ mềm có trọng số thành các cụm và dẫn đến kết quả dày đặc, nhiều thông tin hơn nhưng cũng có nhiều chiều bộ mô tả. Trên thực tế, các tác giả chỉ ra rằng một vectơ Fisher với một hình ảnh duy nhất từ đạt được hiệu suất tương đương với bộ mô tả BoW

với 4.000 từ. Đơn giản hóa với hiệu suất tương đương và đôi khi vượt trội là các vectơ của các bộ mô tả tổng hợp cục bộ (VLAD), được đề xuất bởi J´egou et al. [29]. VLAD vẫn sử dụng cách gán cứng các bộ mô tả cục bộ tới giá trị gần nhất cụm, nhưng nắm bắt phần còn lại của phần tử khôn ngoan của tất cả các tính năng cục bộ từ trung tâm của cụm của họ. Điều đó có nghĩa là vectơ tính năng nhúng được phân vùng thành k phân đoạn, trong đó k là số cụm. Đoạn tương ứng đến trung tâm cụm gần nhất bằng sự khác biệt giữa bộ mô tả cục bộ và trung tâm đó và tất cả các phân đoạn khác là 0. Chiều của quá trình nhúng do đó, không gian là số lượng cụm nhân với kích thước của tính năng cục bộ. Các tổng hợp bao gồm lấy tổng trên tất cả các vectơ đặc trưng cục bộ được chuyển đổi, l2 - bình thường hóa kết quả và áp dụng PCA với làm trắng để giảm độ cao chiều của bộ mô tả toàn cục thành thứ gì đó dễ quản lý hơn (thường là theo thứ tự vài trăm chiều).

Theo định nghĩa, VLAD nhạy cảm với khoảng cách giữa một vectơ đặc trưng cục bộ và trung tâm cụm của nó. Tuy nhiên, khoảng cách Euclide có ý nghĩa hạn chế trong không gian nhiều chiều. Trong một nghiên cứu tiếp theo, J´egou và Zisserman [30] kể lại cho thực tế này bởi L2 -bình thường hóa phần dư, do đó mã hóa góc của chúng thay vì độ lớn của chúng, dẫn đến tên phép nhúng tam giác. Bởi vì khoảng cách không có ý nghĩa, việc gán cứng cho các cụm đơn lẻ là không hợp lý hoặc. Nhúng tam giác do đó mã hóa các góc giữa địa phương vector đặc trưng và tất cả các từ trực quan. Đại diện này sau đó được làm trắng và đã được phát hiện là vượt trội so với vectơ đánh cá và VLAD. Tuy nhiên, Husain và Bober [25] nhận thấy rằng việc so sánh từng vectơ đặc trưng cục bộ với tất cả các từ trực quan không mở rộng thành các tập dữ liệu lớn. Chuyển nhượng cụm mềm, mặt khác, thường cư xử không ổn định và xuống cấp thành một nhiệm vụ duy nhất trong thực tế. Để khắc phục điều này, họ đề xuất một nền tảng trung gian bằng cách chỉ định bộ mô tả cục bộ cho một vài trung tâm

cụm gần nhất và dựa trên trọng số trên hàng ngũ của họ trong số những người hàng xóm gần nhất thay vì khoảng cách thực tế của họ. Hơn nữa, các bộ mô tả trực quan mạnh mẽ (RVD) này không được làm trắng trên toàn cầu nhưng ở cấp độ từng cụm. Các tác giả thấy rằng RVD thực hiện cạnh tranh đến nhúng tam giác, trong khi tính toán nhanh hơn và mạnh mẽ hơn để giảm kích thước.

Vai trò của bộ dữ liệu Trong khi mô hình sử dụng các tính năng cục bộ tổng hợp đối với CBIR bắt đầu từ năm 2003 [48], nghiên cứu trong lĩnh vực này đã được tích cực nhất giữa năm 2010 và 2016. Một lý do có khả năng cho sự chậm trễ này là thiếu sự phù hợp và bộ dữ liệu điểm chuẩn được thiết lập. Trong những năm 2007 và 2008, Tòa nhà Oxford [39], Paris Buildings [40], và INRIA Holidays [28] bộ dữ liệu đã được xuất bản, trong đó nhanh chóng nổi lên như là điểm chuẩn tiêu chuẩn để truy xuất ví dụ và đưa ra động lực cho lĩnh vực này bằng cách cung cấp một cơ sở thích hợp để đánh giá và so sánh của các phương pháp. Hai bộ dữ liệu tòa nhà bao gồm các bức ảnh khác nhau về các địa danh khác nhau các tòa nhà ở Oxford và Paris, với nhiều phối cảnh, quy mô và tắc mạch. Mặt khác, bộ dữ liệu Ngày lễ chứa một tập hợp các ảnh kỳ nghỉ cá nhân với trung bình ba phối cảnh khác nhau cho mỗi cảnh. Mặc dù các bộ dữ liệu này là một thách thức, nhưng nhiệm vụ truy xuất hình ảnh hiển thị cùng một đối tượng hoặc cảnh khi truy vấn được xác định rõ ràng với sự thật rõ ràng. Các tính năng CNN có sẵn Sau khi các tính năng địa phương được chế tạo thủ công vẫn không bị nghi ngờ trong CBIR trong hơn một thập kỷ, sự phục hưng của học sâu cuối cùng đã dẫn đến một sự thay đổi đáng kể liên quan đến biểu diễn hình ảnh. Các công trình độc lập của Babenko et al. [4] và Razavian et al. [45] lần đầu tiên cho thấy rằng có thể đạt được kết quả tốt đáng ngạc nhiên bằng cách đơn giản trích xuất các bộ mô tả hình ảnh toàn cục, được gọi là mã thần kinh, từ lớp được kết nối đầy đủ đầu tiên của CNN sẵn có được đào tạo trước trên ImageNet [14].

Với sự đơn giản cực độ của phương pháp này, đòi hỏi kỹ thuật gần như bằng không nỗ lực so với việc phát hiện các điểm chính, trích xuất các tính năng cục bộ và tổng hợp họ, đây là một kết quả đáng chú ý. Chỉ một năm sau, Babenko và Lempitsky [3] cải thiện đáng kể hiệu suất của phương pháp này bằng cách trích xuất hình ảnh các tính năng không phải từ một kết nối đầy đủ mà từ lớp tích chập cuối cùng, mà vẫn có độ phân giải không gian. Do đó, kết quả là một tập hợp các vectơ đặc trưng, mà đại khái có thể được liên kết với các vùng khác nhau trong hình ảnh. Đây là tổng hợp lên để tổng hợp, L2- được chuẩn hóa, giảm kích thước bằng PCA và l2 - được chuẩn hóa lại, dẫn đến tên gọi tích chập gộp các tính năng (SPoC) cho các bộ mô tả này.

Trong những năm tiếp theo, nghiên cứu chủ yếu tuân theo việc sử dụng trình trích xuất tính năng thần kinh và tập trung vào việc thiết kế tập hợp tinh vi chức năng. Nhiều người trong số họ cố gắng tìm điểm trung gian giữa tổng và cực đại tổng hợp, ví dụ: bằng cách lấy trung bình các lượt kích hoạt trên một số phản hồi hàng đầu chỉ như trong tổng hợp trung bình một phần (PMP) [54] hoặc bằng cách nội suy trơn tru giữa hai cực đoan như trong tổng hợp trung bình tổng quát (GeM) [42]. Tuy nhiên, các tính năng tích chập tổng hợp có một nhược điểm: Ngược lại đối với các tính năng cục bộ truyền thống, chúng không cho phép bản địa hóa chính xác đối tượng phù hợp và do đó, không tương thích với các kỹ thuật như không gian xác minh và xếp hạng lại, phụ thuộc vào thông tin hình học. để này kết thúc, Tolias et al. [50] đề xuất kích hoạt tối đa khu vực tích chập (R-MAC), theo cách tiếp cận hai bước: Tích chập bản đồ tính năng được chia thành các vùng chồng chéo có kích thước khác nhau và cục bộ các vectơ đặc trưng trong mỗi vùng được tổng hợp bằng cách sử dụng tổng hợp tối đa. Những cái này cái gọi là vectơ MAC sau đó được làm trắng và tổng hợp bằng cách tổng hợp thành một bộ mô tả hình ảnh R-MAC toàn cầu. Đối với xếp hạng lại không gian, sự giống nhau của véc-tơ MAC của truy vấn và véc-tơ

MAC khu vực riêng lẻ của số ít hàng đầu kết quả truy xuất có thể được sử dụng để bản địa hóa đối tượng truy vấn trong ảnh được truy xuất

và tinh chỉnh bảng xếp hạng. Các kỹ thuật này đã lấy CBIR dựa trên các tính năng được trích xuất từ được đào tạo trước CNNs khá xa, nhưng bộ mô tả RVD thủ công [25] vẫn có thể cạnh tranh với họ về điểm chuẩn truy xuất phiên bản.

Học từ đầu đến cuối để truy xuất hình ảnh Học sâu cuối cùng đã trở nên vượt trội không thể phủ nhận so với các kỹ thuật CBIR truyền thống dựa trên các tính năng thủ công khi các nhà nghiên cứu bắt đầu điều chỉnh CNN được sử dụng để trích xuất tính năng cho nhiệm vụ truy xuất hình ảnh thay vì sử dụng chương trình được đào tạo trước. Nghiên cứu sinh coi sự thay đổi trọng tâm này từ chuyển đổi và tổng hợp tính năng đến việc học tính năng thực tế như là sự thay đổi mô hình quan trọng thứ hai trong CBIR. Tính năng toàn cầu Gordo et al. [20] là một trong những người đầu tiên thành công trong việc này nỗ lực và thiết lập trạng thái nghệ thuật trong việc truy xuất ví dụ trong ít nhất hai năm. Chúng được xây dựng dựa trên R-MAC [50] và triển khai nó dưới dạng các lớp có thể phân biệt được trên kiến trúc VGG16 CNN, sau đó có thể được đào tạo từ đầu đến cuối. Để đạt được điều này, họ sử dụng triplet loss [47], một mục tiêu đào tạo từ lĩnh vực số liệu sâu học hỏi. Bằng cách đào tạo trên tập dữ liệu được tuyển chọn về các địa danh nổi tiếng, họ học được một đại diện tính năng trong đó hình ảnh của cùng một mốc gần nhau hơn bằng một lề nhất định so với hai hình ảnh của các mốc khác nhau, hỗ trợ mục tiêu truy xuất cá thể.

Cách tiếp cận này sau đó đã được mở rộng bằng cách trích xuất các tính năng R-MAC từ nhiều lớp của một CNN và đánh giá các đặc điểm riêng lẻ của từng khu vực theo Phân kỳ Kullback-Leibler giữa các phân phối của khoảng cách Euclide giữa các mô tả phù hợp và không phù hợp, để phân biệt đối xử tốt hơn các tính năng khu vực có trọng số cao hơn [26]. Động lực để kết hợp các tính năng từ nhiều lớp nằm ở các mức độ trừu tượng trực quan khác nhau: các tính

năng từ các lớp trước đó biểu thị nhiều hơn các thuộc tính trực quan, trong khi các lớp sau các lớp cung cấp một biểu diễn trừu tượng hơn về mặt ngữ nghĩa.

Trái ngược với việc mất bộ ba, Radenovi´c et al. [42] tìm sự mất mát tương phản để cung cấp hiệu suất cuối cùng tốt hơn, trong khi hơn nữa chỉ yêu cầu các cặp thay thế bộ ba hình ảnh để đào tạo. Quan trọng hơn, họ đề xuất một giải pháp không giám sát kỹ thuật tạo dữ liệu huấn luyện bao gồm khớp và không khớp các cặp hình ảnh để truy xuất ví dụ mà không cần chú thích của con người: Hình ảnh trong tập dữ liệu huấn luyện được phân cụm dựa trên biểu diễn BoW của chúng bằng cách sử dụng cục bộ Các tính năng RootSIFT và xác minh không gian được áp dụng để đảm bảo rằng tất cả hình ảnh trong một cụm hiển thị cùng một đối tượng. Một mô hình 3-D sau đó được xây dựng cho từng cụm sử dụng các kỹ thuật cấu trúc từ chuyển động (SfM), để có thể xác định từ những mô hình này cho dù hai hình ảnh mô tả cùng một đối tượng hay không. Điều này cùng CBIR và Lỗ hổng ngữ nghĩa trong kỷ nguyên Deep Learning 9 cho phép hình ảnh của cùng một mốc nhưng được chụp từ các điểm khác nhau và rời rạc quan điểm được coi là không phù hợp. Các thông tin về máy ảnh các vị trí thu được từ SfM hơn nữa cho phép khai thác các vị trí tích cực đầy thách thức các cặp hình ảnh thể hiện số lượng chồng chéo không nhỏ.

Từ việc phân tích những hạn chế của các công trình liên quan ở trên, luận án đề xuất một phương pháp tra cứu ảnh cải tiến hàm khoảng cách, mà cải thiện chức năng khoảng cách dựa trên việc tối đa hóa thương số giữa tổng khoảng cách của các cặp ảnh khác nhau và tổng khoảng cách của các cặp ảnh khác nhau và tổng khoảng cách của các cặp ảnh tương tự. Ở đây luận án xem xét cả tập ảnh tương tự và tập ảnh không tương tự để tìm ma trận trọng số và cải thiện hiệu quả của phương pháp tra cứu.

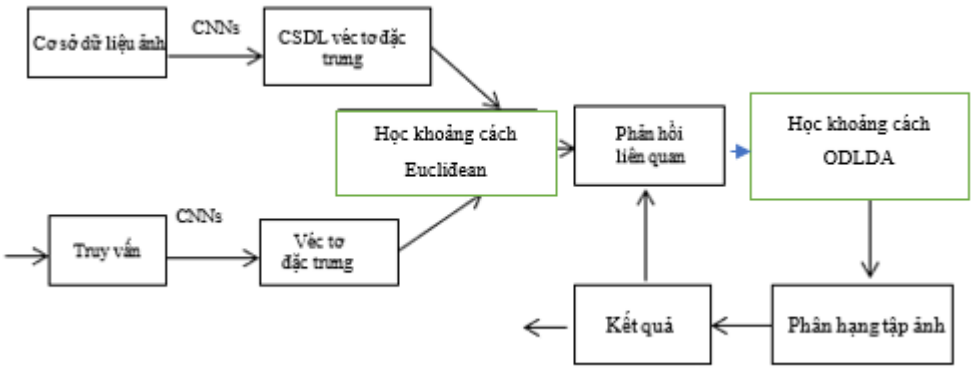
2.3. Đề xuất phương pháp phân hạng lại ảnh

Trong phần này, luận án trình bày ngắn gọn phương pháp đề xuất. Đầu

tiên, phương pháp đề xuất việc xây dựng các đặc trưng sâu để biểu diễn các ảnh. Tiếp theo, trên tập kết quả của pha tra cứu khởi tạo mà sử dụng các đặc trưng sâu, người dùng đánh dấu các ảnh liên quan đến ảnh truy vấn để thu được tập ảnh phản hồi. Tập phản hồi thu được này bao gồm các ảnh liên quan và không liên quan đến ảnh truy vấn. Dựa trên tập ảnh liên quan, phương pháp đề xuất huấn luyện mô hình để tìm phép chiếu tuyến tính. Phép chiếu tuyến tính này thỏa mãn điều kiện mà phương sai giữa các mẫu trong cùng tập liên quan được cực tiểu trong khi cực đại phương sai giữa các mẫu liên quan với các mẫu không liên quan. Bên cạnh đó, phương pháp đề xuất cũng xây dựng một độ đo tương tự Mahalanobis bằng việc tìm ma trận tối ưu M trong công thức độ đo tương tự cải tiến.

2.3.1. Sơ đồ của phương pháp đề xuất

Sơ đồ của phương pháp ODLDA đề xuất được chỉ ra trên Hình II.1. Phương pháp sử dụng mô hình CNN mà được huấn luyện trên tập dữ liệu ImageNet để trích rút đặc trưng sâu. Khi người dùng gửi một ảnh truy vấn, phương pháp trích rút đặc trưng sâu của ảnh truy vấn theo cùng một cách như trích rút đặc trưng của ảnh cơ sở dữ liệu. Sau đó nó so sánh độ tương tự giữa véc tơ đặc trưng của ảnh truy vấn và véc tơ đặc trưng của mỗi ảnh trong cơ sở dữ liệu mà sử dụng khoảng cách Euclide để trả về tập kết quả khởi tạo đối với người dùng. Người dùng thực hiện phản hồi bằng việc đánh dấu các ảnh mà liên quan hoặc không liên quan đối với ảnh truy vấn để thu tập ảnh phản hồi. Sau đó tập ảnh phản hồi được sử dụng như đầu vào đối với thuật toán học độ đo khoảng cách và tối ưu trọng số. Tiếp theo, tất cả các ảnh trong cơ sở dữ liệu ảnh được phân hạng lại, mà dựa vào giá trị của hàm khoảng cách Mahalanobis. Nếu người dùng không thỏa mãn với tập kết quả, quá trình phản hồi sẽ được lặp lại. Nếu người dùng thỏa mãn với kết quả, hệ thống trả về tập kết quả cuối cùng cho người dùng.



Hình II. 3. Sơ đồ của phương pháp đề xuất ODLDA

2.3.2. Tra cứu ảnh sử dụng học sâu

Trong những năm gần đây, mạng CNN đã cho các kết quả tốt trong lĩnh vực thị giác máy tính như phân lớp ảnh, nhận dạng đối tượng, phân đoạn ngữ nghĩa. Trên cơ sở đó, đã có những nghiên cứu về tra cứu ảnh dựa vào nội dung sử dụng CNN và đã thu được các kết quả khả quan.

Trong các tài liệu [4, 21, 61, 62] đã chỉ ra một số cách tiếp cận để cải tiến hiệu quả của hệ thống CBIR sử dụng học sâu để xây dựng tập đặc trưng ngữ nghĩa hơn: 1) mô hình CNN được tiền huấn luyện để xây dựng một tập đặc trưng ảnh với khoảng cách Euclide (chuẩn L_2) để so sánh các độ đo tương tự giữa các véc tơ đặc trưng; 2) nó cũng sử dụng mô hình CNN được tiền huấn luyện để xây dựng tập đặc trưng, nhưng sử dụng thuật toán học độ đo khoảng cách DML để thu độ đo tương tự mà phù hợp hơn với dữ liệu; 3) với tập dữ liệu cụ thể, huấn luyện lại mô hình CNN cùng với một bộ phân lớp cụ thể, sau đó sử dụng độ đo như các cách tiếp cận 1) hoặc 2) để hoàn thành phương pháp tra cứu.

Giả sử NCS có hai ảnh trong cơ sở dữ liệu I_1 và I_2, các đặc trưng sâu được trích rút sử dụng mô hình CNN được tiền huấn luyện trên tập dữ liệu ImageNet.

Đặc trưng sâu của hai ảnh I_1 và I_2 được biểu thị bởi x_1 và x_2. Độ đo tương tự được sử dụng để so sánh hai đặc trưng này là L_2:

$$L_2\_Similarity(x_i, x_j) = \|x_i - x_j\|_2$$

$$= \sqrt{(x_i - x_j)^T (x_i - x_j)} \quad (2.1)$$

Công thức (2.1) chỉ ra độ tương tự giữa các ảnh I_i và I_j, giá trị độ tương tự là lớn hơn cho các ảnh I_i và I_j giống nhau hơn.

Độ đo tương tự sử dụng cách tiếp cận 2) để so sánh hai véc tơ đặc trưng của ảnh được tính bởi công thức L_T:

$$LT\_similarity(x_i, x_j) = \|x_i - x_j\|_T$$

$$= \sqrt{(x_i - x_j)^T T(x_i - x_j)} \quad (2.2)$$

Với một ma trận, thu được từ việc học chỉ số tương tự thỏa mãn điều kiện là ma trận xác định dương, vì chỉ số tương tự phải dương và chỉ số tương tự có giá trị nhỏ nhất khi

$$x_i = x_j$$

Với một ma trận, T thu được từ học độ đo tương tự mà thỏa mãn điều kiện T là một ma trận xác định dương, bởi vì độ đo tương tự phải là dương, và độ đo tương tự có giá trị nhỏ nhất khi $x_1 = x_2$.

Độ đo tương tự ở đây là như trong cách tiếp cận 1) khi ma trận T là một ma trận đơn vị $T = I$. Nói cách khác, nó là một trường hợp đặc biệt khi xét tương quan giữa các thành phần đặc trưng trong cách tiếp cận 1). Hơn nữa, mỗi thành phần đặc trưng có một sự tương tự khác nhau, vậy nó thường là độ đo tương tự trong cách tiếp cận 2) để thu được hiệu quả cao hơn.

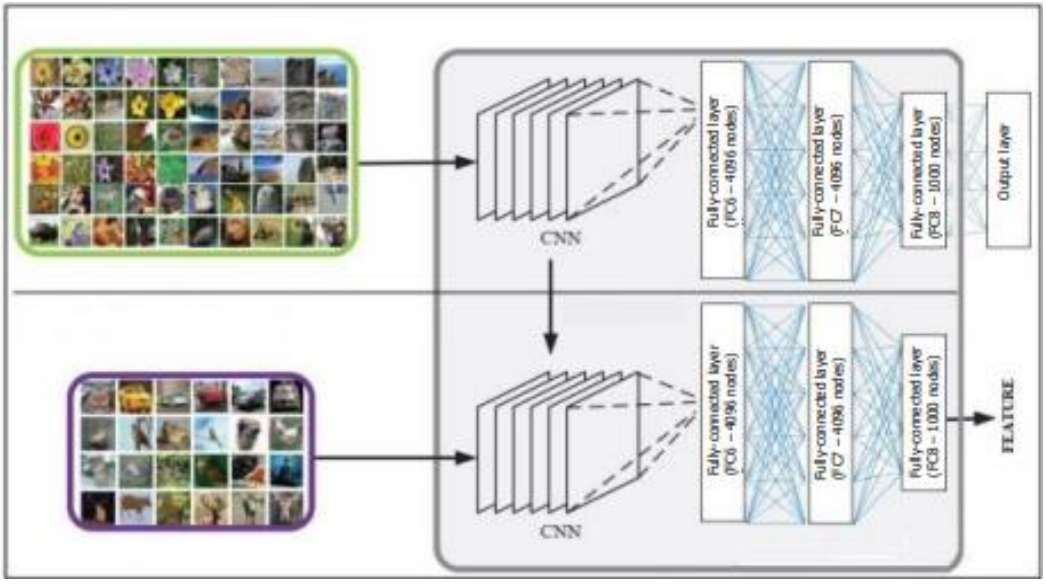
Phương pháp đề xuất thực hiện việc xây dựng các tập đặc trưng dựa vào học sâu. Sau khi thực hiện thủ tục K-NN (K lân cận gần nhất) để thu được một danh sách kết quả khởi tạo và trả chúng về cho người dùng, người dùng sẽ đánh dấu

các ảnh mà liên quan đến ảnh truy vấn để thu được tập phản hồi. Tiếp theo, nó xây dựng một độ đo tương tự cải tiến bằng sử dụng tập mẫu dương, mà lấy cảm hứng từ cách tiếp cận 2). Ma trận M là một ma trận hoàn chỉnh, nó phản ảnh sự tương quan của dữ liệu trên mỗi đặc trưng và giữa các đặc trưng.

Trong phương pháp đề xuất, luận án sử dụng mô hình CNN được tiền huấn luyện trên tập dữ liệu rất lớn. Sau đó nó sử dụng mô hình để trích rút các đặc trưng sâu, cũng được biết như là học biểu diễn ảnh. Lý do chính mà luận án chọn cách tiếp cận này là vì tập dữ liệu đủ lớn cho huấn luyện mô hình không sẵn có. Thêm nữa, để huấn luyện một mô hình CNN, cần rất nhiều thời gian. CNN được sử dụng phổ biến cho phân lớp ảnh, trong đó một ảnh được lan truyền qua mạng và xác suất cuối cùng được lấy từ tầng dưới cùng của mạng. Tuy nhiên, trong quá trình học một biểu diễn, thay vì cho phép ảnh lan truyền trên toàn bộ mạng, có thể dừng tại một tầng bất kỳ, chẳng hạn, tầng kết nối đầy đủ cuối cùng, và trích rút các giá trị từ mạng tại tầng này, sau đó sử dụng chúng như các véc tơ đặc trưng.

Trong phương pháp đề xuất, luận án chỉ sử dụng các tầng tích chập để trích rút các đặc trưng. Mục tiêu là để tổng quát một CNN được tiền huấn luyện trong học các đặc trưng cụ thể của ảnh trong tập dữ liệu. Mô hình được tiền huấn luyện được sử dụng để thu được các véc tơ đặc trưng mạnh hơn một số thuật toán như SIFT, GIST, HOG, vv. Luận án tận dụng khả năng của mô hình mạng nơ ron tích chập CNN, được tiền huấn luyện trong ILSVRC 2012 với 1 , 2 triệu ảnh và 1000 khái niệm để thu các đặc trưng của ảnh. Nó gồm các tầng chập, các lớp gộp (pooling), và các lớp kết nối đầy đủ. Các tầng trước thường là các lớp chập được kết hợp với các hàm kích hoạt phi tuyến và các lớp gộp. Lớp cuối cùng là một lớp kết nối đầy đủ và thường là hàm softmax (Hình II.2). Số các đơn vị trong lớp cuối cùng là 1000. Do đó, đầu ra gần lớp cuối cùng có thể được xem như một véc tơ đặc trưng hữu ích và softmax là bộ phân lớp được

sử dụng. Mô hình sử dụng một đầu vào có cỡ cố định 256*256, trong khi tập dữ liệu được sử dụng trong phương pháp đề xuất có cỡ thay đổi. Do đó, các ảnh được tiền xử lý bởi chuyển chúng sang cỡ 256*256. Khi sử dụng mạng để trích rút đặc trưng cố định, NCS cắt mạng ở một điểm trước tầng kết nối đầy đủ cuối cùng. Do đó, NCS thu được một véc tơ đặc trưng gồm 1000 chiều cho mỗi ảnh.



Hình II. 4. Kiến trúc học biểu diễn dựa vào mô hình CNN được tiền huấn luyện

2.4. Độ đo khoảng cách cải tiến

Cho đến nay, có một số phương pháp học độ đo khoảng cách khác nhau mà tận dụng các thuộc tính của tập phản hồi liên quan trong quá trình tra cứu. Tuy nhiên, các phương pháp đã có chỉ xem xét tập mẫu dương mà bỏ qua tập mẫu âm. Ý tưởng cơ bản của phân tích phân biệt tuyến tính (LDA- linear discriminant analysis) là tìm một biến đổi tối ưu, mà được thực hiện bởi cực đại tổng phương sai giữa các mẫu của các lớp khác nhau (bao gồm lớp âm và lớp dương) và cực tiểu phương sai của dữ liệu trong cũng lớp.

Giả sử rằng tập kết quả tra cứu khởi tạo gồm N ảnh: $X = \{x_i\}_{i=1}^{N}$. Tập kết quả tra cứu khởi tạo được người dùng đánh dấu và trả về cho người dùng, nó cũng được chia thành hai tập tách biệt: một tập mẫu dương và một tập mẫu âm. Để đạt được mục tiêu, cần xác định hai ma trận, S_b và S_w. Ở đây S_b là khoảng cách giữa kỳ vọng *(tâm của lớp)* của các lớp khác nhau và S_w là khoảng cách giữa kỳ vọng và các mẫu của mỗi lớp. Hai ma trận này được tính toán bởi công thức:

$$S_b = \frac{1}{n_b}\sum_{j=1}^{n_b}\sum_{i\in D_j}(m_j - m_i)(m_j - m_i)^T \quad (2.3)$$

$$S_w = \frac{1}{n}\sum_{j=1}^{2}\frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ji} - m_i)(x_{ji} - m_i)^T \quad (2.4)$$

Trong đó n_b là tổng số mẫu của hai tập mẫu dương và âm, m_j là tâm của lớp j, x_{ji} là véc tơ thứ i của lớp j, mỗi D_j là một lớp. Trong bài toán này, có 2 lớp: lớp dương và lớp âm. Tâm m_j của lớp j được tính theo công thức:

$$m_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ji}$$

Thủ tục LDA được gọi là bài toán tối ưu như sau:

$$W = argmax_W \frac{|W'S_bW|}{|W'S_wW|} \quad (2.1)$$

Ma trận W là ma trận biến đổi tối ưu, mà cần phải tìm. Khi thu được biến đổi tối ưu W, nhận trọng số tối ưu của hàm khoảng cách Mahalanobis:

$$W_0 = \mathrm{W^T W}$$

Theo như lý thuyết Fisher [4, 63, 64], bài toán tối ưu (2.5) tương đương với việc cực đại tổng khoảng cách kỳ vọng của các lớp khác nhau $\hat{C}_b$ và cực tiểu tổng khoảng cách kỳ vọng trong cùng lớp S_w [48]. Để tìm nghiệm của bài toán (2.5), luận án đề xuất áp dụng Thuật toán 1.1 ở dưới. Thuật toán này cũng được sử dụng để giải cho các nghiên cứu trước đây về LDA [9].

2.5. Thuật toán tra cứu ảnh

Thuật toán 1.1, gọi là ODLDA, là thuật toán tra cứu ảnh dựa vào phân tích phân biệt tuyến tính và khoảng cách tối ưu.

Algorithm1.1.ODLDA

Input:

Image set: ***DB***

Initialization query image: **Q**

Returned image number for each iteration: ***N***

Output:

Result: ***R***

1. S ←IRL<DB , *M*>;
2. S_q ←IRL<Q , *M*>;
3. $Result_{initial}(Q)$←**$Retrieval_{Initial}$**(Sq , S , N)
4. R← $Result_{initial}(Q)$;
5. **Repeat**

   5.1.< $F_{feature}$, F^{+}_{label}, F^{-}_{label}>)←**Feedback** (R); *relevantfeedback*

   5.2. ***W*** = **LDA(**$F_{feature}$, F^{+}_{label}, F^{-}_{label}**);** *Find the optimal transformation* **W**

   5.3. W_o=W^TW; The *optimal weight of the Mahalanobis distance function*

   5.4. *R*← ***Ranking*** *(S, Wo, N); Rerank the set of images according to the Mahalanobis distance function with the optimal weight.*

   Until (User stops responding);

   $\boldsymbol{Kmeans(f)}$

6**. Return** R ;

Thuật toán ODLDA được thực hiện như sau: Mỗi ảnh trong tập ảnh cơ sở dữ liệu được biểu diễn bởi một véc tơ đặc trưng trong không gian đặc trưng nhiều chiều (Bước 1). Khi người dùng cung cấp một ảnh truy vấn khởi tạo Q, thuật

toán nhúng vào hay trích chọn đặc trưng ảnh truy vấn thành một véc tơ đặc trưng S_q (Bước 2). Truy vấn khởi tạo được thực hiện trong Bước 3 bởi $\text{Result}_{\text{Initial}}(Q) \leftarrow Retrieval_{Initial}\ (S_q, S, N)$, ở đây S_q là biểu diễn của ảnh truy vấn, S là tập biểu diễn của tập ảnh cơ sở dữ liệu và N là số các ảnh được tra cứu trong tập S sau mỗi vòng lặp. Kết quả tra cứu với truy vấn khởi tạo $\text{Result}_{\text{initial}}(Q)$ được gán cho R (Bước 4).

Trên tập $\text{Result}_{\text{initial}}(Q)$ được trả về bởi truy vấn khởi tạo, người dùng phản hồi thông qua hàm Feedback(R) để nhận được tập đặc trưng $F_{feature}$ và tập nhãn $F_{\text{label}}=\{F^{+}_{label}, F^{-}_{label}\}$ (Bước 5.1). Tập phản hồi của người dùng, mà bao gồm các ảnh liên quan và không liên quan, được đưa vào LDA (Bước 5.2) để tìm chiếu A. Tìm chiếu A được thực hiện bởi giải bài toán tối ưu (2.5). Các kết quả của ma trận chiếu này được sử dụng để xây dựng ma trận trọng số tối ưu nhằm cải tiến trọng số của hàm khoảng cách Mahalanobis (Bước 5.3). Ở thời điểm này, thu được hàm khoảng cách Mahalanobis cải tiến như sau:

$$dM(F_i, F_j) = \|F_i - F_j\|_M$$
$$= \sqrt{(F_i - F_j)^T M (F_i - F_j)}$$

Quá trình tra cứu phân lớp lại toàn bộ tập ảnh trong cơ sở dữ liệu bởi hàm *Ranking(S , W_0, N)*, và nhận N ảnh như tập kết quả trả về cho người dùng (Bước 5.4).

2.6. Kết quả thực nghiệm

2.6.1. Môi trường thực nghiệm

1)Tập dữ liệu ảnh Corel: tập ảnh mà luận án sử dụng cho đánh giá thực nghiệm là tập Corel với 10 , 800 ảnh (một số ảnh đại diện như Hình II.3). Một số chủ đề cho tập này gồm onsai, castle, cloud, autumn, aviation, dog, primate, ship, stalactite, fire, tiger, elephant, iceberg, train, waterfall. Mỗi ảnh trong tập

này chứa một đối tượng tiền cảnh. Mỗi chủ đề gồm khoảng 100 ảnh. Cỡ của các ảnh là 120×80 hoặc 80×120.

2) Tập tin cậy nền (Ground truth) cho đánh giá độ chính xác của CBIR: tập tin cậy nền được sử dụng để đánh giá độ chính xác của hệ thống CBIR, tức là, các ảnh liên quan và không liên quan được biết trước ở trong tập tin cậy nền này. Theo đó, hệ thống tra cứu ảnh xem xét các ảnh mà liên quan đến ảnh truy vấn là các ảnh có cùng chủ đề. Tập này gồm ba cột (tiêu đề : Query Image ID, Image ID, and Relation) và bao gồm 1 , 981 , 320 dòng.

3) Tập ảnh SIMPLIcity: Để minh chứng hiệu năng của phương pháp đề xuất, ngoài việc đánh giá thực nghiệm trên tập dữ liệu Corel, luận án cũng thực hiện các thực nghiệm trên tập ảnh SIMPLIcity. Đây là một tập dữ liệu nhỏ với 1000 ảnh và được chia thành 10 chủ đề. Mỗi ảnh trong tập này có cỡ là 256×384 hoặc 384×256. Một số mẫu trong cơ sở dữ liệu ảnh này được chỉ ra trên Hình II.4. Luận án biểu diễn mỗi ảnh bởi hai đặc trưng gồm các đặc trưng màu và các đặc trưng cạnh. Đặc trưng màu được biểu diễn bởi các mô tả cấu trúc màu với một véc tơ 128 chiều, trong khi đặc trưng cạnh là mô tả lược đồ cạnh với véc tơ gồm 150 chiều. Một véc tơ gồm 278 chiều sẽ biểu diễn cho mỗi ảnh. Độ chính xác của phương pháp Baseline (phương pháp không có cơ chế phản hồi) được tính dựa trên khoảng cách Euclide giữa véc tơ đặc trưng 278 chiều của ảnh truy vấn và mỗi ảnh trong cơ sở dữ liệu.

Hình II. 5. Một số mẫu trong thư viện ảnh Corel

Hình II. 6. Một số mẫu trong tập SIMPLIcity

2.6.2. Đánh giá thực nghiệm

Trong thực nghiệm, phương pháp đề xuất được so sánh với năm phương pháp tra cứu ảnh sử dụng các độ đo khoảng cách khác nhau: (1) Euclide; (2) Euclide cải tiến: độ đo Euclide có trọng số của mỗi chiều đặc trưng; (3) Xing: ma trận trọng số và hàm khoảng cách cải tiến, mà là nghiệm của bài toán tối ưu lồi; (4) RCA: độ đo khoảng cách RCA được cải tiến từ khoảng cách Mahalanobis [50, 58, 59]; và (5) MCML: độ đo khoảng cách MCML được cải tiến từ khoảng cách Mahalanobis có tập trọng số là kết quả của biến đổi dữ liệu với các ràng buộc nhãn. Trong thực nghiệm, phương pháp đề xuất ODLDA thực hiện tra cứu trên tập đặc trưng sâu được kết hợp với hàm khoảng cách Mahlanobis. Các kết quả thu được trên ba phạm vi (scope) gồm 50, 100, và

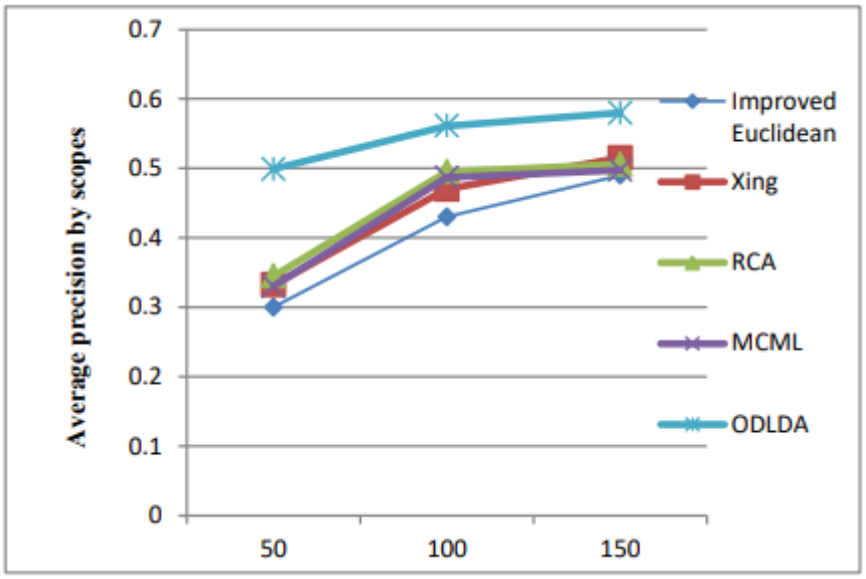
150. Lưu ý rằng giá trị của mỗi scope là top các ảnh được trả về bởi mỗi lần lặp tra cứu. Lý do mà luận án lấy ba scope này là người dùng thường không đủ kiên nhẫn để chọn nhiều hơn 150 phản hồi.

Độ chính xác trung bình của các phương pháp được chỉ ra trên Bảng II.1. Trong bảng này, luận án thấy rằng phương pháp sử dụng độ đo khoảng cách Euclide gốc có độ chính xác thấp nhất. Ba phương pháp Xing, RCA, và MCML có độ chính xác tương tự. Phương pháp đề xuất có độ chính xác cao nhất.

Các đường cong phạm vi – độ chính xác trung bình của Euclide cải tiến, Xing, RCA, MCML và ODLDA được chỉ ra trên Hình II.5. Các giá trị độ chính xác của top 50, 100, và 150 ảnh sau hai lần lặp đầu tiên. Ngoài ra, trên Hình II.5, cũng đưa ra độ chính xác của Baseline cho mục tiêu so sánh. Theo các kết quả này, phương pháp đề xuất thực hiện tốt hơn các phương pháp còn lại. Theo đó, trên hai tập dữ liệu tiêu chuẩn, độ chính xác của phương pháp đề xuất cao hơn của Euclide cải tiến, Xing, RCA, MCML. Điều này khẳng định phương pháp đề xuất là hiệu quả.

Phương pháp	**Độ chính xác trung trình theo các phạm vi (scope)**		
	50	**100**	**150**
Euclide	0.2887	0.3065	0.3199
Euclide cải tiến (Improved Euclidean)	0.3135	0.42658	0.4846
Xing	0.3324	0.47658	0.5125
RCA	0.3424	0.48058	0.5015
MCML	0.3328	0.47958	0.4925
ODLDA	**0.4836**	**0.5065**	**0.5199**

Bảng II. 1. So sánh độ chính xác trung bình của các phương pháp ở scope 50, 100 và 150 trên tập dữ liệu Corel.



Hình II. 7. So sánh độ chính xác trung bình của các phương pháp trên scope 50, 100 và 150 trên tập SIMPLIcity

2.7. Kết luận chương 2

Luận án trình bày phương pháp ODLDA, một kỹ thuật tra cứu ảnh hiệu quả kỹ thuật cải thiện hiệu suất của hệ thống tra cứu ảnh đa điểm. ODLDA khai thác hiệu quả thông tin của người dùng thông qua tập mẫu có liên quan và không liên quan, thực hiện học phép chiếu tối ưu để tách các ảnh không liên quan và thu hẹp khoảng cách của các ảnh liên quan. Phương pháp được đề xuất tìm ma trận trọng số tối ưu của hàm khoảng cách Mahalanobis và sử dụng hàm khoảng cách cải tiến này để xếp hạng toàn bộ tập ảnh cơ sở dữ liệu và trả về tập kết quả cho người dùng. Kết quả thử nghiệm trên hai cơ sở dữ liệu đã chứng minh rằng ODLDA cung cấp độ chính xác cao hơn nhiều so với phương pháp Euclid, Euclid, RCA và OASIS cải tiến.

Kết quả thực nghiệm trên cơ sở dữ liệu đặc trưng gồm 1000 ảnh đã chỉ ra rằng phương pháp được đề xuất cung cấp một độ chính xác cao hơn hẳn so với các phương pháp khác.

Với việc nâng cao hiệu quả tra cứu ảnh dựa vào nội dung bằng cách khoảng cách tối ưu và phân tích phân biệt tuyến tính đã được công bố tại: CT1.

"Improve The Efficiency Of Content-based Image Retrieval Through Incremental Clustering"

Journal of Information Hiding and Multimedia Signal Processing, Vol. 11, No. 3, pp. 103-115, September 2020.

Chương 3. CẢI THIỆN HIỆU QUẢ CỦA TRA CỨU ẢNH DỰA TRÊN NỘI DUNG SỬ DỤNG PHÂN HOẠCH ĐỒ THỊ

Trong những năm gần đây, nhiều phương pháp tra cứu ảnh (CBIR) theo cách tiếp cận phản hồi có liên quan được thiết kế để thu hẹp khoảng trống ngữ nghĩa giữa các đặc trưng trực quan mức thấp và các khái niệm ngữ nghĩa mức cao cho nhiệm vụ tra cứu ảnh. Tuy nhiên, các phương pháp tra cứu ảnh hiện nay chỉ quan tâm đến độ tương tự giữa ảnh truy vấn và ảnh cơ sở dữ liệu mà chưa quan tâm đến độ tương tự giữa các ảnh trong tập ảnh đích. Trong luận án này Nghiên cứu sinh đề xuất một phương pháp tra cứu ảnh hiệu quả sử dụng phân hoạch đồ thị ***(An efficient image retrieval method using a graph clustering-MGC)*** mà khai thác đầy đủ thông tin độ tương tự của tập ảnh. Phần thực nghiệm trên cung cấp các kết quả thực nghiệm để minh chứng độ chính xác của phương pháp đề xuất.

3.1. Nâng cao hiệu quả tra cứu ảnh dựa vào nội dung sử dụng phân hoạch đồ thị

3.1.1. Giới thiệu

Trong xử lý ảnh, đồ thị và phân hoạch đồ thị là các khái niệm quan trọng được sử dụng để mô tả và phân tích các đặc điểm của hình ảnh để cải thiện nâng cao tra cứu ảnh dựa vào nội duug.

*Đồ thị (Graph):

Trong xử lý ảnh, đồ thị thường được sử dụng để biểu diễn mối quan hệ giữa các điểm dữ liệu trên hình ảnh hoặc các thành phần của hình ảnh.

Đồ thị bao gồm các đỉnh (nodes) và các cạnh (edges). Đỉnh thường biểu thị cho các điểm dữ liệu hoặc vị trí trên hình ảnh, trong khi cạnh biểu thị cho mối quan hệ hoặc sự kết nối giữa các điểm đó.

Các đỉnh và cạnh có thể có thông tin bổ sung như trọng số, loại kết nối, hoặc thuộc tính khác, phụ thuộc vào mục tiêu cụ thể của vấn đề xử lý ảnh.

Ví dụ về ứng dụng của đồ thị trong xử lý ảnh bao gồm việc sử dụng đồ thị để biểu diễn các đường viền, mối tương quan giữa các điểm ảnh, hoặc các thành phần kết nối trong hình ảnh.

*Phân hoạch đồ thị (Graph Segmentation):

Phân hoạch đồ thị trong xử lý ảnh là quá trình chia hình ảnh thành các vùng hoặc đối tượng riêng biệt bằng cách sử dụng thông tin từ đồ thị biểu diễn hình ảnh.

Trong phân hoạch đồ thị, các đỉnh trong đồ thị thường biểu thị cho các điểm dữ liệu trên hình ảnh (như điểm ảnh) và các cạnh thể hiện mối quan hệ hoặc sự tương quan giữa các điểm đó.

Mục tiêu của phân hoạch đồ thị là tìm ra các tập con (clusters) của đỉnh trong đồ thị sao cho các điểm trong mỗi tập con thuộc về cùng một vùng trong hình ảnh.

Phân hoạch đồ thị có nhiều ứng dụng trong xử lý ảnh, chẳng hạn như phân đoạn vật thể, tách nền, nhận dạng đối tượng, và nhiều tác vụ khác liên quan đến việc tách biệt và xử lý các phần khác nhau trong hình ảnh.

Phân hoạch đồ thị là một trong các phương pháp tiên tiến trong lĩnh vực xử lý ảnh và thường được sử dụng để giải quyết các vấn đề phức tạp liên quan đến phân vùng và phân tích hình ảnh.

Trong thập kỷ qua, tra cứu ảnh dựa vào nội dung (CBIR) đã thu hút sự quan tâm lớn của cộng đồng nghiên cứu [2 , 3, 4, 36, 38, 39, 57, 65, 66, 67, 68, 69, 70, 71,72,73]. Lựa chọn tính năng được chuẩn hóa bằng đồ thị với việc tái cấu trúc dữ liệu [74] cũng được đặc biệt quan tâm.Trong không gian cao chiều, độ đo khoảng cách Euclide được sử dụng để đo độ tương tự giữa ảnh truy vấn và các ảnh trong cơ sở dữ liệu [4, 20, 21, 22, 23]. Tuy nhiên, do khoảng trống

ngữ nghĩa giữa các đặc trưng mức thấp và các khái niệm ngữ nghĩa mức cao, hệ thống sử dụng độ đo khoảng cách Euclide trong không gian nhiều chiều thường cho hiệu năng nghèo nàn.

Phản hồi liên quan (Relevance feedback - RF) thường được sử dụng để thu hẹp khoảng trống ngữ nghĩa này [3, 7, 9, 23, 48, 75, 76, 77]. RF tập trung vào sự tương tác giữa người dùng và máy tra cứu , nó yêu cầu người dùng gán nhãn các ảnh tương tự hay không tương tự ngữ nghĩa với ảnh truy vấn, trong đó các ảnh tương tự ngữ nghĩa với ảnh truy vấn được xem là các mẫu phản hồi dương và ngược lại là các mẫu phản hồi âm. Một số cách tiếp cận phản hồi liên quan thường được sử dụng trong thập kỷ qua. Máy véc tơ hỗ trợ (support vector machine – SVM) được huấn luyện trên một tập mẫu để thu được một bộ phân lớp nhị phân. Sau đó bộ phân lớp nhị phân (một siêu phẳng tách) được sử dụng để phân hạng các ảnh theo thứ tự giảm dần của khoảng cách đến siêu phẳng tách [4, 9, 19, 20, 47, 75] . Phương pháp của Tao và cộng sự trong [9, 57, 65, 78] đã chia tập mẫu phản hồi âm thành một số tập con và một chuỗi các máy lồi lề nhân được phát triển giữa một nhóm phản hồi dương và một số nhóm phản hồi âm. Việc tích hợp các kỹ thuật học máy phân lớp vào tra cứu ảnh với phản hồi liên quan đã có sự cải tiến hiệu năng, tuy nhiên hiệu năng vẫn cần tiếp tục được nâng cao bởi vì độ đo khoảng cách trong các hệ thống này chưa phù hợp ngữ cảnh vùng cùng bộ.

Nhiều phương pháp tiếp cận mà sử dụng phản hồi liên quan đã được đề xuất gần đây đã nâng cao hiệu năng của hệ thống tra cứu ảnh dựa vào nội dung. Trong các phương pháp này, dựa vào các mẫu phản hồi của người dùng, mô hình học độ đo tương tự được cập nhật hoặc một số tham số đặc trưng được điều chỉnh để phù hợp hơn với mong muốn của người dùng. Từ đó, độ chính xác của phương pháp tra cứu ảnh sẽ được cải thiện dần qua mỗi lần lặp phản hồi. Đào Thị Thúy Quỳnh và cộng sự [65, 78] đã đề xuất một phương pháp tra

cứu ảnh liên quan ngữ nghĩa. Phương pháp giải quyết được các hạn chế: (1) Chỉ sử dụng một truy vấn để tạo ra kết quả tra cứu khởi tạo gồm các ảnh nằm trong các vùng khác nhau; (2) Không thực hiện phân cụm lại tập ảnh phản hồi; (3) xác định được độ quan trọng ngữ nghĩa của từng truy vấn và (4) xác định độ quan trọng theo từng đặc trưng. Những đóng góp này làm cho độ chính xác được cải tiến đáng kể. Nguyễn Hữu Quỳnh và cộng sự [57] nghiên cứu đề xuất phương pháp tra cứu ảnh với hàm khoảng cách thích nghi. Phương pháp đã tận dụng được thông tin địa phương của mỗi vùng điểm trong không gian đặc trưng để xây dựng hàm khoảng cách. Độ chính xác của phương pháp đã được cải tiến đáng kể. Tuy nhiên, độ chính xác của hai phương pháp này còn giới hạn do cả hai phương pháp không xét đến sự không đồng nhất của không gian đặc trưng và không giải quyết vấn đề truy cập xấp xỉ trên không gian non-metric.

Tuy nhiên, các phương pháp tra cứu ảnh sử dụng phản hồi liên quan đề cập ở trên có hạn chế: chỉ quan tâm đến độ tương tự giữa ảnh truy vấn và ảnh cơ sở dữ liệu mà chưa quan tâm đến độ tương tự giữa các ảnh trong tập ảnh đích. Vậy, có thể nâng cao hiệu năng của hệ thống tra cứu ảnh theo cách tiếp cận phản hồi liên quan bằng cách khai thác thông tin tương tự giữa các ảnh trong tập ảnh đích không?

Đây là câu hỏi mà nghiên cứu sinh sẽ giải quyết trong nội dung "Nâng cao hiệu quả tra cứu ảnh dựa vào nội dung sử dụng phân hoạch đồ thị".

3.1.2. Nghiên cứu liên quan:

Giả sử NCS có một cơ sở dữ liệu X gồm n ảnh x_i $(1 \leq i \leq n)$ trong một không gian đặc trưng nhiều chiều R^h, tức là $X = [x_1, x_2, \dots . x_n] \in R^{h \times n}$. Thông tin tiềm nghiệm được cho là các cặp ảnh tương tự S: $(x_i, x_j) \in S$ nếu x_i và x_j được đánh giá là một cặp tương tự, và không tương tự D: $(x_i, x_j) \in D$ nếu x_i và x_j được đánh giá là một cặp không tương tự. Các phương pháp phân tích độ đo khoảng cách nhằm học một độ đo khoảng cách $d_M(x_i, x_j)$ giữa các ảnh x_i và x_j

sao cho các ảnh không tương tự là xa nhau và các ảnh tương tự là gần nhau nhất có thể. Độ đo khoảng cách giữa hai ảnh x_i và x_j được tính như sau:

$$d_M(x_i, x_j) = \|x_i - x_j\|_M = \sqrt{(x_i - x_j)^T M (x_i - x_j)}$$

Ở đây $M \in R^{h \times h}$ là một ma trận nửa xác định dương. Cho $M = I$ có nghĩa là sử dụng độ đo khoảng cách Euclide. Tổng quát hơn, M biểu diễn một học các độ đo khoảng cách Mahalanobis. Bằng việc chấp nhận phân rã giá trị riêng, M có thể được viết lại là $M = AA^T$, $A \in R^{h \times l}$, $l \leq h$.

Học độ đo khoảng cách Mahalanobis M trong không gian đặc trưng nhiều chiều là tương đương với học một ma trận ánh xạ hiệu quả A mà thay thế mỗi ảnh x với $A^T x$ và áp dụng một độ đo khoảng cách Euclide chuẩn vào các ảnh trong không gian thấp chiều.

Các phương pháp phân tích độ đo khoảng cách thường đồng hành với một tập dữ liệu có nhãn với các ràng buộc cặp, chẳng hạn, phân tích thành phần phân biệt (Neighborhood Component Analysis - NCA) được đề xuất để học một độ đo khoảng cách Mahalanobis bằng việc cực tiểu trực tiếp LOOCV (Leave One Out Cross-Validation) của k lân cận gần nhất. Phương pháp lân cận gần nhất lề cực đại (large margin nearest neighbor -LMNN) được đề xuất để đưa lề vào bản miêu tả và tách biệt các mẫu của các lớp khác nhau theo một lề cực đại (Weinberger and Saul, 2009).

Bởi vì cách tiếp cận phân cụm là cơ bản trong phương pháp của nghiên cứu sinh, nghiên cứu sinh sẽ trình bày một số nghiên cứu liên quan về phân cụm trong phần dưới đây.

Các thuật toán phân cụm phân các điểm dữ liệu vào C cụm (hoặc loại) trên cơ sở độ tương tự của chúng. Các thuật toán phân cụm bao gồm phân cụm dựa vào phân hoạch, phân cụm dựa vào mật độ (Rodriguez, Laio, 2014; Khan et al.,

2018), và phân cụm dựa vào đồ thị (Luxburg, 2007; Szlam, Bresson, 2010; Bresson, Laurent, 2013).

Trong các thuật toán phân cụm dựa vào phân hoạch, trung bình của cụm được xem như tâm cụm, và một điểm dữ liệu được gán vào tâm gần nhất. Các thuật toán phân cụm dựa vào mật độ, các cụm là các nhóm điểm dữ liệu được đặc trưng bởi cùng mật độ cục bộ, và một tâm cụm là điểm dữ liệu có mật độ cao nhất. Các thuật toán phân cụm dựa vào đồ thị xác định một đồ thị với các đỉnh bằng với các thành của một tập dữ liệu, và các cạnh được đánh trọng số bởi độ tương tự giữa các cặp điểm dữ liệu trong tập dữ liệu. Sau đó các thuật toán tìm một phân hoạch tối ưu của đồ thị sao cho các cạnh giữa các đồ thị con khác nhau có một trọng số rất thấp và các cạnh trong cùng một đồ thị con có trọng số cao. Có một số cấu trúc phổ biến để biến đổi một tập dữ liệu sang một đồ thị tương tự, như đồ thị k lân cận gần nhất (KNN) (Luxburg, 2007). Các tiêu chuẩn cắt đồ thị được sử dụng phổ biến bao gồm cắt tối thiểu (min cut), cắt theo tỷ lệ (ratio cut), cắt chuẩn hóa (normalized cut-Ncut).

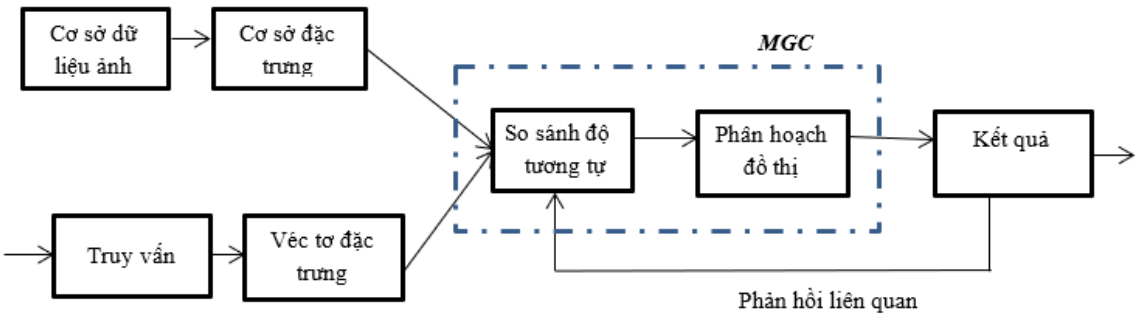
Phân cụm các tập dữ liệu tách được phi tuyến là một vấn đề thách thức trong phân tích phân cụm. Nhiều phương pháp đã được đề xuất để giải quyết vấn đề này. Phương pháp nhân ánh xạ một tập dữ liệu tách được phi tuyến vào một không gian Hilbert cao chiều. Trong không gian này, tập dữ liệu có thể tách đươc tuyến tính. Phân cụm DBK (Marin et al., 2019) đề xuất một nguyên lý cân bằng mật độ. Dựa vào nguyên lý này, họ đề xuất một thuật toán phân cụm nhân thích nghi. Nhiều thuật toán phân cụm nhân (Jia et al, 2016; Yu et al., 2012) sử dụng nhiều hàm nhân để tăng cường hiệu năng của các thuật toán phân cụm nhân. Các thuật toán K-mean nhân với các hàm nhân thích hợp có thể phân cụm các tập dữ liệu tách được phi tuyến, nhưng nó khó để lựa chọn các hàm nhân thích hợp.

Phân cụm phổ là một thuật toán phân cụm dựa vào độ thị nổi tiếng. Đầu tiên, nó xây dựng một ma trận Laplacian đồ thị, sau đó nó tính toán các giá trị riêng và các véc tơ riêng của ma trận Laplacian đồ thị. Nó đề cập đến các véc tơ riêng tương ứng với k giá trị riêng nhỏ nhất như các nhúng chiều thấp của tập dữ liệu, và cuối cùng sử dụng một số thuật toán phân cụm cơ bản (chẳng hạn, K-mean) để thu được kết quả phân cụm. Phương pháp phân cụm siêu phẳng (Hofmeyr, 2017) thiết lập một khung siêu phẳng để giải quyết bài toán Ncut.

Phân cụm không gian con thưa (Elhamifar, Vidal, 2013) xây dựng một đồ thị tương tự bởi các kỹ thuật biểu diễn thưa, và sau đó sử dụng phân cụm phổ để tính toán các kết quả phân cụm. Phân cụm cung cấp các kết quả phân cụm tốt cho các tập dữ liệu tách được phi tuyến, nhưng nó là phức tạp để tính các giá trị riêng và các véc tơ riêng.

3.1.3. Phương pháp đề xuất:

Phương pháp ***MGC*** được mô tả bởi lược đồ trên Hình III.1. Với một truy vấn mà người dùng đưa vào, phương pháp sẽ thực hiện trích rút đặc trưng của ảnh truy vấn. Sau đó, sẽ so sánh độ tương tự giữa ảnh truy vấn và cơ sở đặc trưng sử dụng hàm khoảng cách Euclid để có tập kết quả khởi tạo. Sau khi có được tập kết quả khởi tạo với N mẫu, phương pháp thực hiện phân hoạch N ảnh mẫu này thành k cụm. Đến đây, phương pháp sẽ hiển thị k cụm ảnh trả về cho người dùng. Nếu người dùng đã thỏa mãn với tập hết quả này thì đó là kết quả cuối cùng. Trong trường hợp người dùng chưa thỏa mãn thì quá trình phản hồi liên quan sẽ được thực hiện, tập ảnh liên quan do người dùng lựa chọn sẽ được đưa vào phân hoạch và trả về kết quả cập nhật cho người dùng. Quá trình được lặp lại cho đến khi người dùng ngừng phản hồi và phương pháp đưa ra tập kết quả.



Hình III. 1. Sơ đồ của tra cứu ảnh sử dụng phân hoạch đồ thị

3.1.4. Phân cụm cắt tối thiểu lặp (Iterative Min Cut Clustering)

Trong lĩnh vực khoa học máy tính, một ứng dụng quan trọng của phân hoạch đồ thị là phân cụm dữ liệu sử dụng một mô hình đồ thị - các độ tương tự cặp giữa tất cả các đối tượng dữ liệu tạo ra một ma trận kề đồ thị có trọng số mà chứa tất cả thông tin cần thiết cho phân cụm. Bài toán phân hoạch đồ thị sử dụng trọng số "Cut" của một tập dữ liệu S = $\{s_1, \ldots, s_n\} \subset R^d$ thành C cụm bằng cách xây dựng một đồ thị và tra cứu một phân vùng đồ thị sao cho các đỉnh trong cùng một cụm thì càng tương tự và các đỉnh ở khác cụm thì càng khác xa nhau. Bài toán này phân hoạch đồ thị sử dụng trọng số "Cut" là một bài toán NP_khó, có nhiều phương pháp giải xấp xỉ khác nhau, trong số phân cụm quang phổ (speactral clustering) được sử dụng khác rộng rãi.

Iterative Min Cut Clustering

Phương pháp Iterative Min Cut Clustering (IMC) được đề xuất phân cùng một tập dữ liệu X = $\{x_1, \ldots, x_N\} \subset R^H$ thành C cụm bằng cách tối thiểu hóa hàm mục tiêu:

$\sum_{i,j} w_{ij}$, x_i và x_j thuộc các cụm khác nhau (3.1)

với w_{ij} là độ tương đồng (trọng số cạnh) giữa x_i và x_j. Để việc tính toán cho thuận tiện, ta chuẩn hóa các điểm dữ liệu x_i (i $\epsilon \{1, \ldots, N\}$) như sau:

$$x_i = \frac{x_i}{max\{x_i[1],\ldots,x_N[H]\}} \quad (3.2)$$

Độ tương tự w_{ij} được tính bằng:

$$w_{ij} = \begin{cases} exp(-\frac{||x_i - x_j||^2}{2\sigma^2}) \text{ , } x_i \text{ và } x_j \text{ là các láng giềng} \\ 0 \text{ nếu ngược lại} \end{cases}$$

–(3.3)

Để giải quyết vấn đề **(1)**, ta định nghĩa một feature q (là đại lượng vô hướng) cho mỗi điểm dữ liệu. Nếu 2 điểm dữ liệu thuộc cùng một cụm thì q của chúng sẽ có giá trị giống nhau và ngược lại. Có q_i đại diện cho feature của x_i , $q_i = q_j$ nếu x_i và x_j thuộc cùng một cụm và $q_i \neq q_j$ nếu ngược lại. véc tơ $q = [q_i] = [q_1, \ldots, q_N]^T$ có thể được xem như một chiều được gán của tập dữ liệu X. **(1)** tương đương với:

$$Q = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} (q_i - q_j)^2 \quad (3.4)$$

Dựa vào mối quan hệ giữa **(4)** và ma trận Laplacian:

$$q^T L q = \frac{1}{2} \sum_{i,j} w_{ij} (q_i - q_j)^2 \quad (3.5)$$

Để giải quyết vấn đề ***(3.4)***:

$$\frac{\partial Q}{\partial q_i} = 2 \sum_j (q_i - q_j) w_{ij} - 2j \sum_j (q_i - q_j) w_{ji} = 4 \sum_j (q_i - q_j) w_{ij} \quad (3.6)$$

$$\frac{\partial Q}{\partial q_i} = 0 \Rightarrow q_i = \frac{\sum_j w_{ij} q_i}{\sum_j w_{ij}} \quad (3.7)$$

Theo phương pháp biến phân thì f chứa 2 giá trị của f, có thể được coi như f^k và f^{k+1}.

Khi có được véc tơ đặc trưng f rồi, ta phân vùng cho véc tơ f thành C cụm bằng cách sử dụng một số thuật toán cơ bản như K-means hoặc dùng phương pháp ngưỡng như sau:

$$L_i = \begin{cases} 0 \ nếu \ f_i < T_1 \\ .... \\ c \ nếu \ T_c < f_i < T_{c+1} \\ ..... \\ C \ nếu \ f_i > T_C \end{cases}$$

Với T_c là ngưỡng thứ c.

Từ đó, ta có thuật toán IMC giải quyết vấn đề ***(3.4)*** như sau :

Thuật toán phân cụm IMC

Input: X

Output: c cụm: T_1, T_2 , … , T_C

Tính w_{ij} theo công thức (3.3), khởi tạo ngẫu nhiên cho

Lặp:

Tính f^{n+1} với $f_i^{(n+1)} = \frac{\sum_j w_{ij} f_j^{(n)}}{\sum_j w_{ij}}$

Cho đến khi $|\boldsymbol{f^{(n)} - f^{(n+1)}}|$ nhỏ hơn một dung sai quy định hoặc n đã đạt số vòng lặp tối đa.

Return T_1, T_2 , … , T_C

Thuật toán tra cứu

Thuật toán 1.3 dưới đây là mô tả thuật toán tra cứu ảnh hiệu quả sử dụng phân hoạch đồ thị (An efficient image retrieval method using a graph clustering-MGC)

Thuật toán 1.3. Thuật toán tra cứu ảnh *MGC*

Input:

Tập các ảnh: ***S***

Ảnh truy vấn: $\boldsymbol{Q_{initial}}$

Số các ảnh được trả về tại mỗi lần lặp: N

Output:

Danh sách kết quả tổng hợp: $\boldsymbol{Result(Q_{merger})}$

1. Result($Q_{initial}$) ←< q, *d*, *S*, *N*>;

3. IMC (Result($Q_{initial}$, N), C, X)

5. **Repeat**

5.1 **for** i=1 to C do

Result(Q_{merger}) ← <{$q^{(1)}$, $q^{(2)}$, ..., $q^{(c)}$}, *d*, *S*, *N*>;

5.3 Relevant(Q_{merger} , M)←***Feedback***(Result(Q_{merger}) , N');

until (User dừng phản hồi);

6. Return Result(Q_{merger});

Thuật toán tra cứu ảnh hiệu quả sử dụng phân hoạch đồ thị (An efficient image retrieval method using a graph clustering-MGC) thực hiện như sau:

Đầu tiên, người dùng đưa vào truy vấn Q, thuật toán hàm khoảng cách Euclidean d trên không gian và trả về N ảnh thông qua < q, *d*, *S*, *N*> để được các kết quả của truy vấn khởi tạo Result ($Q_{initial}$). Tiếp theo, thuật toán sẽ thực hiện phân hoạch trên tập kết quả khởi tạo Result ($Q_{initial}$) sử dụng thuật toán IMC và trả C cụm ảnh. Sau đó, thuật toán sẽ thực hiện tìm đại diện cho mỗi cụm. Dựa trên C đại diện mỗi cụm tương ứng với C truy vấn và hàm khoảng cách d, thuật toán trả về N' ảnh kết quả trên tập ảnh S thông qua < {$q^{(1)}$, $q^{(2)}$, ..., $q^{(c)}$}, *d*, *S*, *N*> và gán cho Result (Q_{merger}). Trên tập kết quả Result (Q_{merger}), người dùng chọn M ảnh liên quan thông qua hàm ***Feedback***(Result(Q_{merger}) , N') để

có tập Relevant(Q_{merger} , M). Quá trình này được lặp đi lặp lại cho đến khi người dùng dừng phản hồi. Kết thúc thuật toán đưa ra một tập các ảnh kết quả Result (Q_{merger}).

3.2. Thực nghiệm

3.2.1. Môi trường thực nghiệm

Để xác định hiệu quả của các mô hình và phương pháp đề xuất, thực nghiệm được xây dựng trên nền tảng dotNET, ngôn ngữ lập trình C#, Python và Matlab. Cấu hình máy tính sử dụng để thực nghiệm: Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz, DDRam - 16GB và hệ điều hành Windows 11 Professional.

Thực nghiệm được mô tả dưới hai dạng gồm: đồ thị và bảng biểu; trong đó, hiệu suất tra cứu về độ chính xác và phạm vi được mô tả bằng đồ thị, các bảng biểu mô tả chỉ số đánh giá trung bình và so sánh giữa các phương pháp với nhau.

CSDL ảnh thực nghiệm SIMPLIcity

Để chứng minh hiệu quả của phương pháp đề xuất, nghiên cứu sinh tiến hành thử nghiệm trên Dataset SIMPLIcity. Đây là một tập ảnh gồm 1.000 ảnh với 10 chủ đề. Mỗi ảnh có ích thước 256×384 hoặc 384×256. Một số ảnh được đưa ra trong Hình III.2. Nghiên cứu sinh tực hiện biểu diễn mỗi ảnh bởi hai đặc trưng: đặc trưng mầu sắc và đặc trưng kết cấu. Đặc trưng mầu được biểu diễn bởi véc tơ 128 chiều, đặc trưng cạnh được biểu diễn bởi véc tơ 150 chiều. Mỗ ảnh sẽ được biểu diễn bởi mộtt véc tơ 278 chiều thể hiện đặc trưng mầu sắc và kết cấu của ảnh trong cơ sở dữ liệu SIMPLIcity.

Hình III. 2. Một số ảnh trong tập SIMPLIcity

3.2.2. Thực hiện truy vấn và đánh giá

Trong phần thực nghiệm, các tham số được lựa chọn như sau:

Hiệu quả tra cứu được đánh giá trên cơ sở dữ liệu ảnh **SIMPLIcity** gồm 1000 ảnh, tất cả các ảnh trong cơ sở dữ liệu được sử dụng để thực hiện các truy vấn. Thực nghiệm thực hiện đánh giá độ chính xác của phương pháp đề xuất dựa trên độ chính xác trung bình của 1.000 ảnh truy vấn. Mỗi truy vấn thực hiện sẽ trả về 500 ảnh. Nhằm mục đích đánh giá, nghiên cứu sinh sử dụng độ chính xác trung bình để đánh giá hiệu quả và so sánh với ba phương pháp khác. Độ chính xác trung bình là tỷ lệ của số ảnh liên quan trong danh sách trả về cho người dùng và được tính toán bởi trung bình tất cả các truy vấn. Độ lệch chuẩn dùng để đo lường độ biến thiên của độ chính xác trung bình.

So sánh độ chính xác trung bình của phương pháp đề xuất

Thực nghiệm thực hiện 1.000 truy vấn dưới 3 cấu hình (2, 4, 6 cụm) để nhận được độ chính xác trung bình. Ba phương pháp khác nhau được sử dụng

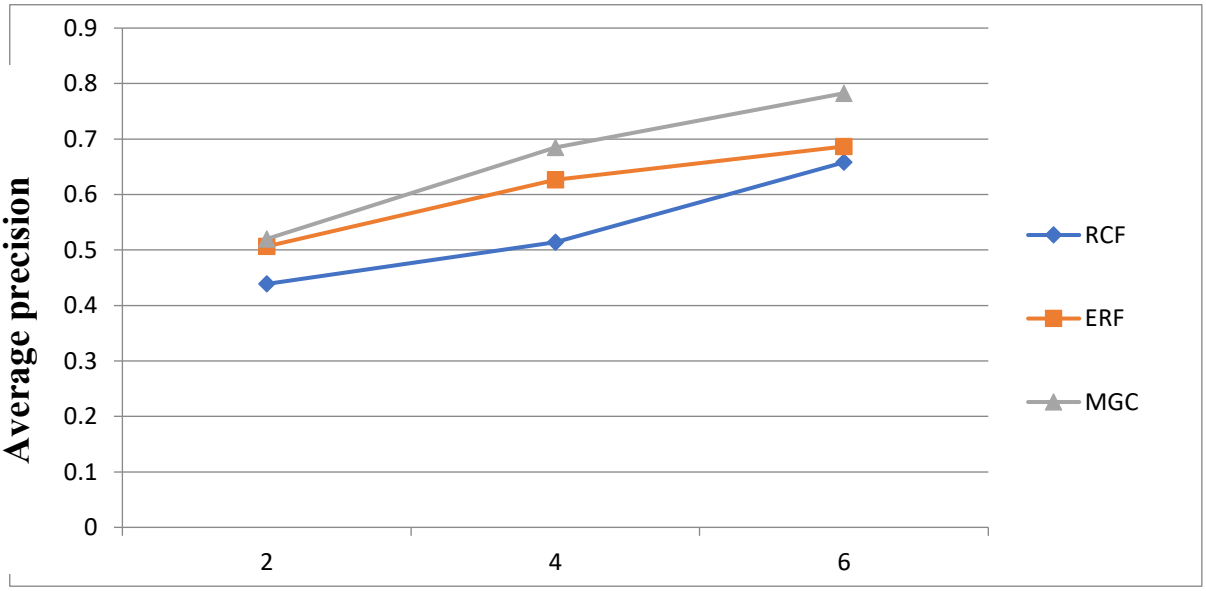
để so sánh bao gồm **CRF** (Complementary Relevance Feedback) [48], **ERF** (Efficient Relevance Feedback) [79] và phương pháp đề xuất **MGC**. Hơn nữa có 3 lần lặp phản hồi được dùng trong thực nghiệm đánh giá. Các kết quả thực nghiệm được chỉ ra trong Hình III.3. Trục ngang chỉ ra số điểm truy vấn là 2, 4 và 6 điểm truy vấn tương ứng với số đại diện của mỗi cụm. Trục đứng chỉ ra độ chính xác. Năm phương pháp khác nhau gồm **CRF**, **ERF** và **MGC**.

.

Phương pháp	**Trung bình độ chính xác**		
	2	**4**	**6**
CRF	0.4388	0.5065	0.5199
ERF	0.5138	0.62658	0.6846
MGC	**0.658**	**0.68658**	**0.7825**

Bảng III. 1. Bảng kết quả trung bình độ chính xác của 3 phương pháp theo số điểm truy vấn trong ba lần phản hồi.

Trong Bảng III.1. thể hiện độ chính xác trung bình của ba phương pháp là CRF, ERF, và phương pháp đề xuất **MGC** tại các mức 2, 4 và 6 điểm truy vấn, với phương pháp đề xuất số điểm truy vấn được xác định theo số cụm. Với 2 điểm truy vấn, độ chính xác của phương pháp đề xuất cao hơn hai phương pháp CRF, ERF là 12.92%, 21.92%. Trường hợp 4 điểm truy vấn, độ chính xác của phương pháp đề xuất CRF, ERF là 12.00%, 6%. Trường hợp 8 điểm truy vấn, phương pháp đề xuất có độ chính xác cao hơn CRF, ERF lần lượt 16.47%, 26.26%.



Hình III. 3. So sánh độ chính xác của ba phương pháp trên tập ảnh SIMPLIcity

Kết luận chương 3

Ngoài ra nghiên cứu sinh đã tập trung vào đề xuất phương pháp, có tên thuật toán tra cứu ảnh hiệu quả sử dụng phân hoạch đồ thị (An efficient image retrieval method using a graph clustering-MGC) mà khai thác đầy đủ thông tin độ tương tự của tập ảnh. Kết quả thực nghiệm của nghiên cứu sinh trên cơ sở dữ liệu đặc trưng ảnh đã chỉ ra rằng phương pháp được đề xuất ***MGC*** cung cấp một độ chính xác cao hơn hẳn so với các phương pháp khác.

Với việc nâng cao hiệu quả tra cứu ảnh dựa vào nội dung bằng phân hoạch đồ thị đã được công bố tại:

CT2. Phuong Nguyen Thi Lan, Tao Ngo Quoc, Quynh Dao Thi Thuy, Minh-Huong Ngo, ***"Improve the Effectiveness of Image Retrieval by Combining the Optimal Distance and Linear Tách Analysis"*** International Journal of Advanced Computer Science and Applications, https://dx.doi.org/10.14569/IJACSA.2021.0120206, 2021.

CT3. The – Anh Pham1, Dinh – Nghiep Le, ***"PRODUCT SUB-VECTOR QUANTIZATION FOR FEATURE INDEXING"*** Journal of Computer Science and Cybernetics, V.35, N.1 (2019), 69–83 DOI 10.15625/1813-9663/35/1/13442.

CT4. Nguyễn Thị Lan phương, Đào Thị Thúy Quỳnh, Ngô Quốc Tạo, Nguyễn Ngọc Quỳnh, Lê Hưng, "*Nâng cao hiệu quả tra cứu ảnh dựa vào nội dung sử dụng phân hoạch đồ thị* ", Hội thảo quốc gia lần thứ XXV, Một số vấn đề chọn lọc của công nghệ thông tin và truyền thông. Hà Nội, ngày 8-9/12/2023. Nhà xuất bản Khoa học Kỹ thuật, trang 129-134 , 2022.

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Luận án trình bày phương pháp ODLDA, một kỹ thuật tra cứu ảnh hiệu quả kỹ thuật cải thiện hiệu suất của hệ thống tra cứu ảnh đa điểm. ODLDA khai thác hiệu quả thông tin của người dùng thông qua tập mẫu có liên quan và không liên quan, thực hiện học phép chiếu tối ưu để tách các ảnh không liên quan và thu hẹp khoảng cách của các ảnh liên quan. Phương pháp được đề xuất tìm ma trận trọng số tối ưu của hàm khoảng cách Mahalanobis và sử dụng hàm khoảng cách cải tiến này để xếp hạng toàn bộ tập ảnh cơ sở dữ liệu và trả về tập kết quả cho người dùng. Kết quả thử nghiệm trên hai cơ sở dữ liệu đã chứng minh rằng ODLDA cung cấp độ chính xác cao hơn nhiều so với phương pháp Euclid, Euclid, RCA và OASIS cải tiến.

Ngoài ra nghiên cứu sinh đã tập trung vào đề xuất phương pháp, có tên thuật toán tra cứu ảnh hiệu quả sử dụng phân hoạch đồ thị (An efficient image retrieval method using a graph clustering-MGC) mà khai thác đầy đủ thông tin độ tương tự của tập ảnh. Kết quả thực nghiệm của nghiên cứu sinh trên cơ sở dữ liệu đặc trưng ảnh đã chỉ ra rằng phương pháp được đề xuất ***MGC*** cung cấp một độ chính xác cao hơn hẳn so với các phương pháp khác.

Tóm lại, luận án đã đạt một số kết quả như sau:

Thứ nhất là: Cải tiến phương pháp tra cứu ảnh thông qua tìm một phép đo khoảng cách tối ưu, mà giảm khoảng cách giữa các cặp ảnh có độ tương tự cao và tối đa hóa khoảng cách giữa các cặp ảnh có độ tương tự thấp.

Thứ hai là: Đề xuất phương pháp tra cứu ảnh dựa trên lý thuyết cắt đồ thị, mà không phải tính ma trận Laplacian, các giá trị riêng và các véc tơ riêng.

Tuy nhiên, luận án còn một số hạn chế: phương pháp giải quyết trong luận án mới được đánh giá trên cơ sở dữ liệu vừa mà chưa xem xét trên các cơ sở dữ liệu lớn.

Từ những hạn chế trên hướng nghiên cứu tiếp theo của luận án là: tích hợp với mô hình học sâu để thích hợp với cơ sở dữ liệu lớn và tăng độ chính xác

DANH MỤC CÁC CÔNG TRÌNH CỦA LUẬN ÁN

TẠP CHÍ KHOA HỌC

CT1. Quynh Dao Thi Thuy, Quynh Nguyen Huu, Phuong Nguyen Thi Lan, Tao Ngo Quoc, Minh-Huong Ngo, ***"Improve The Efficiency Of Content-based Image Retrieval Through Incremental Clustering"*** Journal of Information Hiding and Multimedia Signal Processing, Vol. 11, No. 3, pp. 103-115, September 2020.

CT2. Phuong Nguyen Thi Lan, Tao Ngo Quoc, Quynh Dao Thi Thuy, Minh-Huong Ngo, ***"Improve the Effectiveness of Image Retrieval by Combining the Optimal Distance and Linear Tách Analysis"*** International Journal of Advanced Computer Science and Applications, https://dx.doi.org/10.14569/IJACSA.2021.0120206, 2021.

CT3. The – Anh Pham1, Dinh – Nghiep Le, Phuong Nguyen Thi Lan, ***"PRODUCT SUB-VECTOR QUANTIZATION FOR FEATURE INDEXING"*** Journal of Computer Science and Cybernetics, V.35, N.1 (2019), 69–83 DOI 10.15625/1813-9663/35/1/13442.

CT4. Nguyễn Thị Lan Phương, Đào Thị Thúy Quỳnh, Ngô Quốc Tạo, Nguyễn Ngọc Quỳnh, Lê Phú Hưng, "*Nâng cao hiệu quả tra cứu ảnh dựa vào nội dung sử dụng phân hoạch đồ thị* ", Hội thảo quốc gia lần thứ XXV, Một số vấn đề chọn lọc của công nghệ thông tin và truyền thông. Hà Nội, ngày 8-9/12/2023. Nhà xuất bản Khoa học Kỹ thuật, trang 129-134 , 2022.

DANH MỤC CÁC CÔNG TRÌNH LIÊN QUAN

CT5. Hà Mạnh Toàn, Nguyễn Văn Năng, Trịnh Hiền Anh, Nguyễn Thị Lan Phương, ***"Một số kỹ thuật phân lớp người sử dụng mạng nơron tích chập",*** Hội thảo quốc gia lần thứ XXI, Một số vấn đề chọn lọc của công nghệ thông tin và truyền thông.

CT6. Trần Sơn Hải, Lê Quang Thái, Lê Hoàng Thái, Ngô Quốc Tạo, ***"Phương pháp kết hợp TLD và CMT cho theo vết đối tượng chuyển động",*** Hội thảo quốc gia lần thứ XXI, Một số vấn đề chọn lọc của công nghệ thông tin và truyền thông.

CT7. Jeng-Shyang Pan, Truong-Giang Ngo, Thi-Kien Dao, Thi-Thanh-Tan Nguyen, Trong-The Nguyen, ***"An Optimizing Cross-Entropy Thresholding for Image Segmentation based on Improved Cockroach Colony Optimization"*** JIHMSP, Vol.11, No.4, 2020.

CT8 Nguyễn Thị Lan Phương, Đỗ Văn Hải, Hoàng Văn Hùng, Trần Phạm Văn Cương, ***"Xây dựng cơ sở dữ liệu tổng hợp phục vụ phát triển du lịch tỉnh Lào Cai bằng công nghệ GIS và viễn thám"*** Tạp chí Khoa học và Công nghệ Đại học Thái Nguyên. Tập 225, số 07/1, 2020

TÀI LIỆU THAM KHẢO

[1] M. Flickner, *Query by image and video content: The QBIC system*, Computer vol. 28, no. 9, pp. 23-32, 1995.

[2] [Wayne NiblackRonald J BarberRonald J BarberWilliam Equitz], "*The QBIC Project: Querying Images by Content, Using Color, Texture, and Shape*", Proceedings of SPIE - The International Society for Optical Engineering 1908:173-187, January 1993

[3] D. Liu, K. A. Hua, K. Vu, and N. Yu, (2009) *"Fast Query Point Movement Techniques for Large CBIR Systems"*, IEEE Transactions on Knowledge and Data Engineering, vol. 21, No. 5, pp. 729-743.

[4] A. Alzu'bi, A. Amira, & N. Ramzan, (2017, August). Content-based image retrieval with compact deep convolutional features. Neurocomputing, 249, 95-105.

[5] S.P. Rana, M. Dey, & P. Siarry, (2019, January). Boosting contentbased image retrieval performance through integration of parametric & nonparametric approaches. Journal of Visual Communication and Image Representation, 58(3), 205-219.

[6] Xiao, Z., & Qi, X. "*Complementary relevance feedback-based content-based image retrieval*". Multimedia Tools Appl., 2014, 73(3), 2157–2177.

[7] L. Zhang, H. P. H. Shum, and L. Shao, *"Discriminative Semantic Subspace Analysis for Relevance Feedback"*, IEEE Trans. Image Processing, Vol. 25, No. 3, pp. 1275–1287, Mar. 2016.

[8] M. Yousuf, Z. Mehmood, H.A. Habib, T. Mahmood, T. Saba, A. Rehman, & M. Rashid, (2018). A novel technique based on visual words fusion analysis of sparse features for effective content-based image retrieval. Mathematical Problems in Engineering, 2018, 1-13.

[9] D. Tao, X. Tang, X. Li, and X. Wu, *"Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval,"* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 7, pp. 1088 –1099, July 2006.

[10] A. Raza, H. Dawood, H. Dawood, S. Shabbir, R. Mehboob, & A. Banjar (2018). Correlated primary visual texton histogram features for content base image retrieval. IEEE Access, 6, 46595-46616.

[11] Ishikawa, Y., Subramanya, R., Faloutsos, C., 1998. Mind Reader: *"Querying databases through multiple examples"*. In: Proceedings of the 24th VLDBConference, New York, USA, pp. 218–227.

[12] Jin, X., & French, J.C, (2005), *"Improving Image Retrieval Effectiveness via Multiple Queries"*, Multimedia Tools and Applications, vol. 26, pp. 221-245.

[13] Chengjun Liu and Guangwei Song Management Department, Shenzhen University, Shenzhen, *"A Method of Measuring the Semantic Gap in Image Retrieval: Using the Information Theory"* China, 978 – 1 61284-881-5/11/$26.00©2011 IEEE.

[14] E. Xing, A. Ng, and M. Jordan. "Distance metric learning with application to clustering with side-information". In NIPS, 2002.

[15] Wu, L., Faloutsos, C., Sycara, K., Payne, T.R., (2000). FALCON: *"Feedback adaptive loop for content-based retrieval"*. In: Proceedings of the 26th VLDB Conference, Cairo, Egypt, pp.297–306.

[16] M. Guillaumin, J. J. Verbeek, and C. Schmid. *"Is that you? metric learning approaches for face identification"*. In ICCV, pages 498–505, 2009.

[17] H. Bay, A. Ess, T. Tuytelaars, & L. Van Gool, (2008, June). *Speeded-Up Robust Features (SURF)*. Computer Vision and Image Understanding, 110(3), 346-359.

[18] A. Khoder and F. Dornaika. *A hybrid discriminant embedding with feature selection: application to image categorization.* Applied Intelligence, pages 1-17, 2020.

[19] Andre B, Vercauteren T, Buchner AM, Wallace MB, Ayache N (2012). *"Learning semantic and visual similarity for endomicroscopy video retrieval"* IEEE Transactions on Medical Imaging. 31(6):1276–88.

[20] M.J.J. Ghrabat, G. Ma, I.Y. Maolood, S.S. Alresheedi, & Z.A. Abduljabbar, (2019, December). An effective image retrieval based on optimized genetic algorithm utilized a novel SVM-based convolutional neural network classifier. Human-centric Computing and Information Sciences, 9(1), 31.

[21] A. Voulodimos, N. Doulamis, A. Doulamis, & E. Protopapadakis, (2018). Deep learning for computer vision: A brief review. Computational Intelligence and Neuroscience, 2018, 1-13.

[22] I. Gogul, & V.S. Kumar, "Flower species recognition system using convolution neural networks and transfer learning," in 2017 Fourth International Conference on Signal Processing, Communication and Networking (ICSCN), 2017, no. March, pp. 1-6. Chennai, India.

[23] M. Tzelepi, & A. Tefas, (2018, January). Deep convolutional learning for content based image retrieval. Neurocomputing, 275, 2467–2478.

[24] N. Shrivastava, & V. Tyagi, (2015, August). An efficient technique for retrieval of color images in large databases. Computers & Electrical Engineering, 46, 314-327.

[25] Vũ Văn Hiệu, *Nghiên cứu một số kỹ thuật phân hạng trong tra cứu ảnh dựa vào nội dung*, Học viện Khoa học và Công nghệ-Viện Hàn lâm KH&CN Việt Nam, (Luận án tiến sĩ, 2017).

[26] Y. Weiss, *"Segmentation using eigenvectors: a unifying view"*, in Proc. Int. Conf. Computer Vision, 1999, pp. 975–982.

[27] J. Shi and J. Malik, *"Normalized cuts and image segmentation*", IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 888–905, Aug. 2000.

[28] Y. Weiss, *"Segmentation using eigenvectors: a unifying view"* inProc. Int. Conf. Computer Vision, 1999, pp. 975–982.

[29] J. **Wen, X. Fang, J. Cui, L. Fei, K. Yan, Y. Chen, and Y. Xu**. Robust sparse linear discriminant analysis. IEEE Transactions on Circuits and Systems for Video Technology, 29(2):390-403, 2018.

[30] A. S. Mian, Y. Hu, R. Hartley, and R. A. Owens. *"Image set based face recognition using self-regularized non-negative coding and adaptive distance metric learning"*. IEEE Transactions on Image Processing, 22(12):5252–5262, 2013.

[31] Monique Laurent, Franz Rendl, *"Semidefinite Programming and Integer Programming"*, Report PNA-R0210, CWI, Amsterdam, April 2002.

[32] Sathiamoorthy, S., Natarajan, M. (2020), An efficient content-based image retrieval using enhanced multi-trend structure descriptor. SN Appl. Sci. 2, 217.

[33] Jia, L.; Li, M.; Zhang, P.; Wu, Y.; Zhu, H. *"SAR Image Change Detection Based on Multiple Kernel K-Means Clustering with Local-Neighborhood Information"*. IEEE Geosci. Remote Sens. Lett. 2016, 13, 856–860. [CrossRef].

[34] M. Sajjad, A. Ullah, J. Ahmad, N. Abbas, S. Rho, & S.W. Baik, (2018, February). Integrating salient colors with rotational invariant texture

features for image representation in retrieval systems. Multimedia Tools and Applications, 77(4), 4769-4789.

[35] N. Tadi Bani, & S. Fekri-Ershad, (2019, August). Contentbased image retrieval based on combination of texture and colour information extracted in spatial and frequency domains. Electronic Library, 37(4), 650-666.

[36] M.K. Alsmadi, (2020, April). Content-based image retrieval using color, shape and texture descriptors and features. Arabian Journal for Science and Engineering, 45(4), 3317-3330.

[37] [Eakins J P (1993)] *"Design criteria for a shape retrieval system"* *Computers in Industry* **21**, 167-184.

[38] A. Ponomarev, H.S. Nalamwar, I. Babakov, C.S. Parkhi, & G. Buddhawar, (2016, February). Content-based image retrieval using color, texture and shape features. Key Engineering Materials, 685, 872-876.

[39] Z. Zhao, Q. Tian, H. Sun, X. Jin, & J. Guo, (2016, January). Content based image retrieval scheme using color, texture and shape features. International Journal of Signal Processing Image Processing Pattern Recognition, 9(1), 203-212.

[40] P. Srivastava, & A. Khare, (2017, January). Integration of wavelet transform, local binary patterns and moments for content-based image retrieval. Journal of Visual Communication and Image Representation, 42, 78-103.

[41] H. Lu and R. Mazumder. Randomized gradient boosting machine. SIAM Journal on Optimization, 30(4):2780-2808, 2020.

[42] Porkaew, K., Chakrabarti, K., (1999). *"Query refinement for multimedia similarity retrieval in MARS"*. In: Proceedings of the 7th ACM Multimedia Conference, Orlando, Florida, pp. 235–238.

[43] A. Y. Ng, M. I. Jordan, and Y. Weiss. *"On spectral clustering: Analysis and algorithm"*. In Proceedings Of Neural Information Processing Systems (NIPS), 2001.

[44] Rodriguez, A.; Laio, A. *"Clustering by Fast Search and Find of Density Peaks"*. Science 2014, 344, 1492–1496. [CrossRef] [PubMed]

[45] Hongbo Luo, Sujuan Zhou, *"Image Retrieval of Poisonous Mushrooms Based on Relevance Feedback and Clustering Algorithm"*, Proceedings of the Second Intefrnational Conference on Mechatronics and Automatic Control pp 685-694.

[46] -Yu S.; Tranchevent, L.; Liu, X.; Glanzel, W.; Suykens, J.A.; De Moor, B.; Moreau, Y. "Optimized Data Fusion for Kernel k-Means Clustering". IEEE Trans. Pattern Anal. Mach. Intell. 2012, 34, 1031–1039. [PubMed].

[47] R. Wang, X. Wang, S. Kwong, C. Xu (2017), Incorporating diversity and informativeness in multiple-instance active learning. IEEE Trans Fuzzy Syst 25(6):1460-1475.

[48] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N.Yianilos(2000). *"The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments"*. IEEE Transactions on Image Processing, 9(1):20–37.

[49] Jiebo Luoa, Andreas E. Savakis, Amit Singhal, *"A Bayesian network-based framework for semantic image understanding"* Pattern Recognition 38 (2005) 919 – 934.

[50] Z. Wang, Y. Hu, and L.-T. Chia. *"Learning image-to-class distance metric for image classification"*. ACM TIST, 4(2):34, 2013.

[51] D. W. Jacobs, D. Weinshall, and Y. Gdalyahu, *"Classification with nonmetric distances: image retrieval and class representation,"* IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 6, pp. 583–600, Jun. 2000.

[52] C. Domeniconi, J. Peng, and D. Gunopulos. *"Locally adaptive metric nearest-neighbor classification"*. IEEE Trans. Pattern Anal. Mach. Intell., 24(9):1281–1285, 2002.

[53] A. Sezavar, H. Farsi, & S. Mohamadzadeh, (2019, August). Content-based image retrieval by combining convolutional neural networks and sparse representation. Multimedia Tools and Applications, 78(15), 20895-20912.

[54] A. Patil, & M. Rane, (2021). Convolutional neural networks: an overview and its applications in pattern recognition. In Smart Innovation, Systems and Technologies, 195(Insights into Imaging), 21-30.

[55] K. Simonyan, & A. Zisserman, (2014, September). Very deep convolutional networks for large-scale image recognition. 3rd International Conference Learning Represent ICLR 2015 - Conference Track Proceedings, 1-14.

[56] D. H. Wolpert. Stacked generalization. Neural Networks, 5(2):241-259, 1992.

[57] (Quynh et al., 2018) Quynh Nguyen Huu, Quynh Dao Thi Thuy, Canh Phuong Van, Can Nguyen Van, Tao Ngo Quoc (2018), "*An efficient image retrieval method using adaptive weights",* Applied Intelligence, Volume 48, pp 3807–3826.

[58] G. McLachlan. Discriminant Analysis and Statistical Pattern Recognition. John Wiley, 1992.

[59] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. *"Large scale online learning of image similarity through ranking"*. Journal of Machine Learning Research, 11:1109–1135, 2010.

[60] Hameed, I.M., S.H. Abdulhussain, and B.M. Mahmmod (2021), *Content-based image retrieval: A review of recent trends.* Cogent Engineering, 2021. **8**(1): p. 1927469.

[61] J. Wan, D. Wang, S. C. H. Hoi, and et al, *"Deep learning for contentbased image retrieval: A comprehensive study, "* ACM International Conference on Multimedia, pp. 157-166, 2014.

[62] U. Mittal, S. Srivastava, & P. Chawla, "Review of different techniques for object detection using deep learning," in Proceedings of the Third International Conference on Advanced Informatics for Computing Research - ICAICR '19, New York, USA, 2019, pp. 1-8.

[63] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller. *"Fisher discriminant analysis with kernels"*. In Proc. IEEE NN for Signal Processing Workshop, pages 41–48, 1999.

[64] Q. Liu, H. Lu, and S. Ma. *"Improving kernel fisher discriminant analysis for face recognition"*. IEEE Trans. on Circuits and Systems for Video Technology, 14(1):42–49, 2004.

[65] Quynh Dao Thi Thuy, Quynh Nguyen Huu, Canh Phuong Van, Tao Ngo Quoc (2017), *"An efficient semantic – Related image retrieval method"*, Expert Systems with Applications, Volume 72, pp. 30-41.

[66] G.A. Montazer, & D. Giveki, (2015, September). Content based image retrieval system using clustered scale invariant feature transforms. Optik (Stuttg), 126(18), 1695-1699.

[67] A. Babenko, A. Slesarev, A. Chigorin, and et aIl, *"Neural codes for image retrieval,* " vol. 868, pp. 584-599, 2014.

[68] Nguyễn Thị Uyên Nhi, Văn Thế Thành, Lê Mạnh Thạnh, *"Nâng cao hiệu quả truy vấn hình ảnh theo ngữ nghĩa trên cây phân cụm C-Tree"*. Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 09-10/8/2018 DOI: 10.15625/vap.2018.00049

[69] Văn Thế Thành, Lê Mạnh Thạnh. *"Tra cứu ảnh theo nội dung dựa trên chỉ mục mô tả đặc trưng trong thị giác"* Kỷ yếu kỷ niệm 35 năm thành lập Trường ĐH Công nghiệp Thành Phố Hồ Chí Minh -2017).

[70] Ja-Hwung Su; Wei-Jyun Huang; Vincent S. Tseng, *"Efficient Relevance Feedback for Content-Based Image Retrieval by Mining User Navigation Patterns"*, IEEE Transactions on Knowledge and Data Engineering Volume: 23 Issue: 3, 360-372.

[71] Q. Zheng, X. Tian, M. Yang, & H. Wang, (2019). Differential learning: A powerful tool for interactive content-based image retrieval. Engineering Letters, 27(1), 202-215.

[72] R. Ashraf, M. Ahmedm, S. Jabbar, S. Khalid, A. Ahmad, S. Din, & G. Jeon, (2018, March). Content based image retrieval by using color descriptor and discrete Wavelet transform. Journal of Medical Systems, 42 (3), 44.

[73] B.S. Phadikar, A. Phadikar, & G.K. Maity, (2018, May). Content-based image retrieval in DCT compressed domain with MPEG-7 edge descriptor and genetic algorithm. Pattern Analysis and Applications, 21(2), 469-489.

[74] Z. Zhao, X. He, D. Cai, L. Zhang, W. Ng, and Y. Zhuang. Graph regularized feature selection with data reconstruction. IEEE Transactions on Knowledge and Data Engineering, 28(3):689-700, 2015.

[75] P. Hong, Q. Tian, and T. S. Huang, *"Incorporate support vector machines to content-based image retrieval with relevance feedback,"* in Proceedings of the IEEE International Conference on Image Processing, 2000, pp. 750 –753.

[76] Rui, Y., Huang, T., Mehrotra, S., 1997. *"Content-based image retrieval with relevance feedback in MARS"*. In: Proceedings of IEEE International Conference on Image Processing, October, Santa Barbara, CA.

[77] X. S. Zhou and T. S. Huang, *"Relevance feedback in image retrieval: A comprehensive review,"* Multimedia Systems, vol. 8, no. 6, pp. 536–544, Apr. 2003.

[78] D. T T Quynh, N H Quynh, PV Canh, NQ Tao, *"An efficient semantic – Related image retrieval method"*, Expert Systems with Applications, Volume 72, pp. 30-41, 2017.

[79] K. Kira and L. A. Rendell. A practical approach to feature selection. In Machine learning proceedings 1992, pages 249–256. Elsevier, 1992.

[80] Bai, J. Chen, L. Huang, K. Kpalma, & S. Chen, (2018, January). *Saliency-based multi-feature modeling for semantic image retrieval*. Journal of Visual Communication and Image Representation, 50, 199-204.

[81] J.-E. Lee, R. Jin, and A. K. Jain. *"Rank-based distance metric learning: An application to image retrieval"*. In CVPR, 2008.

[82] -Marin, D.; Tang, M.; Ayed, I.B.; Boykov, Y. *"Kernel Clustering: Density Biases and Solutions"*. IEEE Trans. Pattern Anal. Mach. Intell. 2019, 41, 136–147. [CrossRef] [PubMed]

[83] Elhamifar, E.; Vidal, R. *"Sparse Subspace Clustering: Algorithm, Theory, and Applications"*. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 35, 2765–2781.

[84] Y. Chen, J.Z. Wang, and R. Krovetz, (2003). *"An unsupervised learning approach to content-based image retrieval"*, Seventh International Symposium on Signal Processing and its Applications (ISSPA 2003), Paris.

[85] J. Costeira and T. Kanade, *"A multibody factorization method for motion analysis,"* inProc. Int. Conf. Computer Vision, 1995, pp. 1071–1076.

[86] M. Garg, & G. Dhiman, (2020, June). A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants. Neural Computing & Applications, 33(4), 1311-1328.

[87] S. Chaudhry, & R. Chandra, (2016, October). Unconstrained face detection from a mobile source using convolutional neural networks. Lecture Notes in Computer Science, 9948, 567-576. Including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics.

[88] K. He, X. Zhang, S. Ren, & J. Sun, (2016). Deep Residual Learning for Image Recognition. Computer Vision and Pattern Recognition (pp.770-778). IEEE.

[89] L. Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.

[90] L. Deng and J. C. Platt. Ensemble deep learning for speech recognition. In Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[91] M. J. Laan. van der, eric c. polley, and alan e. hubbard. "super learner". Statistical Applications in Genetics and Molecular Biology, 6, 2007.

[92] L.K. Pavithra, & T. Sree Sharmila, (2019, December). An efficient seed points selection approach in dominant color descriptors (DCD). Cluster Computing, 22(4), 1225-1240.

[93] **C.E. Jacobs, A. Finkelstein, & D.H. Salesin**, "Fast multiresolution image querying," in Proceedings of the 22nd annual conference on Computer graphics and interactive techniques - SIGGRAPH '95, New York, USA, 1995, pp. 277-286.