

DẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

SINH VIÊN THỰC HIỆN:
NGUYỄN VĂN BIÊN - 10520245
PHẠM DUY - 10520074

KHOÁ LUẬN TỐT NGHIỆP

**NGHIÊN CỨU KỸ THUẬT
VÀ XÂY DỰNG ỨNG DỤNG
TÌM KIẾM ĐỐI TƯỢNG TRÊN ẢNH**



LUẬN VĂN CỦA NHÂN KHOA HỌC MÁY TÍNH

GIẢNG VIÊN HƯỚNG DẪN:
TS. NGÔ ĐỨC THÀNH
PGS. TS. LÊ ĐÌNH DUY

6, 2014

LỜI CÁM ƠN

Tôi xin chân thành cảm ơn ...

TÓM TẮT

Trong những năm gần đây, truy vấn ảnh trên tập dữ liệu lớn là bài toán đang thu hút được nhiều sự quan tâm và có ý nghĩa quan trọng trong thực tiễn. Bài toán trên có thể phát biểu như sau: Dựa vào một hình ảnh có chứa đối tượng quan tâm và ngay lập tức trả về những hình ảnh có chứa đối tượng đó từ một tập dữ liệu trong thời gian thực. Các hệ thống truy vấn ảnh trên cơ sở dữ liệu lớn có nhiều ứng dụng quan trọng trong các lĩnh vực như nhận dạng đối tượng hay địa điểm, tìm kiếm video, phát hiện trùng lặp và tái tạo 3D, v.v... Tuy nhiên, bài toán trên cũng đang đối mặt với nhiều thách thức. Bên cạnh vấn đề về sự xuất hiện các biến thể của hình ảnh của đối tượng do sự khác nhau về độ sáng, kích thước, góc chụp hay bị che khuất một phần thì ở đây còn một vấn đề quan trọng khác là phải đảm bảo được thời gian thực hiện truy vấn đặc biệt là khi tìm kiếm trong tập dữ liệu lớn.

Rất nhiều công trình nghiên cứu đã được đề xuất để giải quyết vấn đề trên và đã đạt được nhiều bước tiến đáng chú ý. Hầu hết các công trình đó đều dựa trên mô hình Bag-of-Words (BoW), theo đó mỗi hình ảnh sẽ được biểu diễn bằng các đặc trưng cục bộ, sau đó các đặc trưng này được lượng tử hóa vào các visual word. Để tăng hiệu suất của quá trình truy vấn, người ta thường sử dụng mô hình Bag-of-Words kết hợp với phương pháp đánh chỉ mục ngược (Inverted Index). Thế nhưng cả Bag-of-Words và Inverted Index đều bỏ qua một thông tin quan trọng để tăng độ chính xác cho truy vấn, đó là thông tin không gian ảnh (spatial information) của các đặc trưng cục bộ.

Trong luận văn này, chúng tôi đề xuất một phương pháp nhằm tích hợp thông tin không gian ảnh vào phương pháp đánh chỉ mục ngược (Inverted Index) để nâng cao độ chính xác nhưng vẫn đảm bảo được

thời gian truy vấn nhanh. Kết quả thí nghiệm trên các tập dữ liệu chuẩn như Oxford 5k, Paris 6k và Holiday đã cho thấy tính hiệu quả của phương pháp này.

Từ khóa: Tìm kiếm ảnh - Image Search, Kích cỡ lớn - Large-Scale, Thông tin không gian - Spatial Information, Chỉ mục ngược - Inverted Index.

Mục lục

Mục lục	iv
Danh sách hình vẽ	vi
Danh sách bảng	vii
Danh sách từ viết tắt	viii
1 Tổng quan	1
1.1 Đặt vấn đề	1
1.1.1 Một vài hướng ứng dụng của hệ thống truy vấn ảnh	2
1.2 Thách thức	4
1.3 Mục tiêu và phạm vi của đề tài	5
1.4 Cấu trúc luận văn	6
2 Các công trình liên quan	7
2.1 Biểu diễn hình ảnh bằng các đặc trưng cục bộ	8
2.2 Mô hình Bag-of-words	9
2.2.1 Truy vấn văn bản	10
2.2.2 Bag-of-words trong truy vấn ảnh	11
2.3 Sử dụng thông tin không gian ảnh trong truy vấn ảnh	13
2.3.1 Các hướng tiếp cận dựa trên đặc trưng hình học	14
2.3.2 Các hướng tiếp cận dựa trên thông tin không gian của các đặc trưng cục bộ	15
2.4 Kết chương	15

MỤC LỤC

3 Phương pháp đề xuất	16
3.1 Chỉ mục ngược với biểu diễn Bag-of-Visual-Words	16
3.2 Tích hợp thông tin không gian ảnh vào chỉ mục ngược	18
4 Thực nghiệm và đánh giá kết quả	22
4.1 Các bộ dữ liệu và phương thức đánh giá	22
4.1.1 Các bộ dữ liệu	22
4.1.1.1 Oxford 5k	22
4.1.1.2 Paris 6k	23
4.1.1.3 Holidays	23
4.1.2 Phương thức đánh giá	25
4.2 Cài đặt thí nghiệm	26
4.3 Kết quả thí nghiệm và đánh giá kết quả	27
5 Kết luận và hướng phát triển	28
5.1 Kết luận	28
5.2 Hướng phát triển	29
Phụ lục A	30
Phụ lục B	31
References	32

Danh sách hình vẽ

1.1	Sơ đồ tổng quát của một hệ thống truy vấn ảnh	2
1.2	Những thay đổi bè ngoài của đối tượng trên ảnh	4
2.1	Các từ trực quan (visual words)	12
2.2	Bỏ qua thông tin không gian ảnh trong mô hình bag-of-words .	13
3.1	Quá trình tạo tập tin chỉ mục ngược	17
3.2	Khái quát về phương pháp đề xuất	20
3.3	Quá trình truy vấn của phương pháp đề xuất	21
4.1	Landmark và các truy vấn được dùng để đánh giá	24

Danh sách bảng

Danh mục từ viết tắt

BoW Bag-of-Words

SPM Spatial Pyramid Matching

mAP mean Average Precision

Chương 1

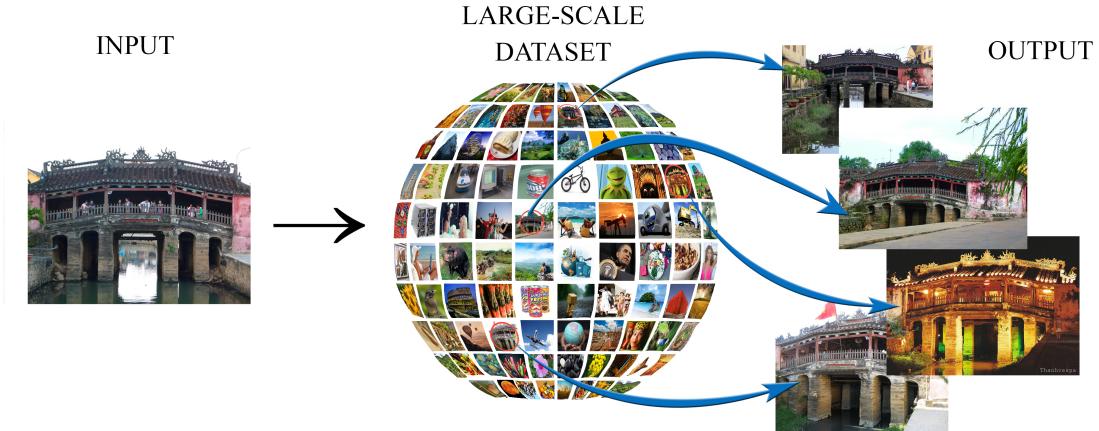
Tổng quan

1.1 Đặt vấn đề

Trong những năm gần đây, cùng với sự phát triển của công nghệ thông tin, các lĩnh vực liên quan đến kỹ thuật số cũng đang có tốc độ phát triển chóng mặt. Các thiết bị kỹ thuật số như máy ảnh, máy quay phim kỹ thuật số, camera số, điện thoại di động có chức năng chụp hình, ... đang ngày càng phổ biến và không ngừng gia tăng về số lượng. Chính điều này đã làm sản sinh ra một lượng thông tin số khổng lồ bao gồm hình ảnh, video, v.v... Do đó, nhu cầu truy vấn thông tin từ kho dữ liệu hình ảnh, video ngày càng bức thiết hơn bao giờ hết.

Để đáp ứng yêu cầu đó, rất nhiều hệ thống truy vấn ảnh đã ra đời. Với đầu vào là một tấm hình có chứa đối tượng quan tâm, hệ thống sẽ trả về những hình ảnh hoặc video từ kho dữ liệu có sẵn mà có chứa đối tượng đó. Hình ảnh [1.1](#) minh họa tổng quát cho một hệ thống truy vấn đối tượng trên ảnh.

Những hệ thống truy vấn ảnh trên tập dữ liệu lớn có rất nhiều ứng dụng trong thực tế. Từ những ứng dụng phục vụ nhu cầu truy vấn thông tin hàng ngày cho tới những ứng dụng giúp quản lý kho dữ liệu lớn trong doanh nghiệp hay dùng để hỗ trợ cho các hệ thống khác. Chúng tôi sẽ liệt kê sơ lược một vài ứng dụng của hệ thống này trong mục dưới đây.



Hình 1.1: Sơ đồ tổng quát của một hệ thống truy vấn ảnh

1.1.1 Một vài hướng ứng dụng của hệ thống truy vấn ảnh

Trong cuộc sống, ta có thể dễ dàng bắt gặp những ứng dụng vô cùng hữu ích của các hệ thống truy vấn đối tượng trên ảnh. Dưới đây là một vài hướng ứng dụng cụ thể:

Nhận dạng đối tượng, sản phẩm. Với sự phổ biến của điện thoại thông minh và internet, một người có thể dễ dàng dùng điện thoại chụp một tấm hình và hỏi hệ thống về thông tin của đối tượng trong tấm hình đó. Ví dụ, tại một cửa hàng, một người mua hàng có thể tham khảo giá của một sản phẩm tại các cửa hàng khác; trong thư viện, một độc giả có thể tìm được những cuốn sách nào chứa hình ảnh mình quan tâm; khi đi thăm bảo tàng, du khách có thể tìm kiếm thêm thông tin về một hiện vật trong đó, v.v...

Nhận dạng địa điểm. Vị trí địa lý của nơi chụp tấm hình cũng có thể được xác định bằng việc truy vấn thông tin của đối tượng trong hình từ những cơ sở dữ liệu lớn chứa hình ảnh và thông tin vị trí như Google Street View hay kho hình ảnh có lưu kèm thông tin GPS. Hệ thống này có thể là một giải pháp thay thế rẻ tiền cho các thiết bị có GPS. Chẳng hạn, khi một du khách đến một nơi mà anh ta chưa bao giờ đặt chân tới nhưng lại không GPS hay bản đồ, anh ta có thể chụp một tấm hình của một tòa nhà hay những cảnh tại nơi đó để xác định được vị trí chính xác của mình.

Tìm kiếm và quản lý kho dữ liệu video. Hàng ngày, một lượng lớn dữ liệu video được sinh ra và ta không thể nào quản lý hết được nội dung của chúng. Ví dụ, một đài truyền hình muốn tìm kiếm tất cả các đoạn quảng cáo có liên quan đến một nhãn hiệu sản phẩm mà họ đã từng phát trong vài năm gần đây, một hệ thống truy vấn ảnh sẽ dễ dàng thực hiện điều này chỉ với một hình ảnh của sản phẩm.

Gán nhãn ảnh tự động. Những tấm ảnh có thể được gán nhãn một cách tự động về địa điểm hay đối tượng trong hình để dễ dàng cho việc tìm kiếm và quản lý sau này. Ví dụ, người dùng có thể dễ dàng tìm kiếm được những bức hình chụp tại một địa điểm nào đó mà không cần biết nó nằm trong album nào hay được chụp ngày nào. Những hệ thống lớn lưu trữ ảnh lớn như của Facebook có thể dễ dàng phát hiện và gán nhãn khuôn mặt người nhưng vẫn chưa thể nhận dạng được địa điểm mà tấm hình được chụp từ nội dung chứa trong hình.

Sử dụng trong quảng cáo theo ngữ cảnh. Rất nhiều công ty quảng cáo đặt màn hình tại nơi công cộng để quảng cáo cho các sản phẩm của mình nhưng các quảng cáo này chưa thực sự hướng người dùng và kém hiệu quả. Việc sử dụng một hệ thống có thể quảng cáo theo ngữ cảnh và hướng đúng đối tượng người dùng sẽ giúp việc quảng cáo hiệu quả hơn. Ví dụ, một camera trong thang máy có thể tự động phát hiện được những sản phẩm người đi thang máy đang dùng như nhãn hiệu chai nước họ đang uống, nhãn hiệu quần áo họ đang mặc,... để lựa chọn được những quảng cáo phù hợp với đối tượng người dùng và phát trên màn hình.

Tăng tính tương tác thực tế. Với sự ra đời của các sản phẩm công nghệ gần gũi với cuộc sống như Google Glass, việc nhận dạng đối tượng trong thời gian thực sẽ mang đến nhiều thông tin hữu ích cho người dùng.

Hỗ trợ cho các hệ thống thị giác máy tính khác. Hệ thống truy vấn đối tượng có thể được dùng để hỗ trợ cho các hệ thống thị giác máy tính khác. Một ví dụ điển hình là hệ thống tự động tái tạo hình ảnh ba chiều sẽ cần gom cụm các hình ảnh của cùng một đối tượng từ một tập dữ liệu lớn.

1.2 Thách thức

Để giải quyết bài toán truy vấn đối tượng trên tập dữ liệu ảnh lớn, có rất nhiều thách thức được đặt ra. Dưới đây chúng tôi sẽ trình bày một vài thách thức trong bài toán này:

Sự biến đổi bề ngoài của đối tượng trong hình ảnh. Một hệ thống truy



Hình 1.2: **Những thay đổi bề ngoài của đối tượng trên ảnh.** (i) Hình ảnh đối tượng trong các điều kiện chiếu sáng khác nhau. (ii) Hình ảnh đối tượng dưới các góc chụp khác nhau. (iii) Đối tượng bị che khuất hay hình ảnh đối tượng bị cắt ghép. (iv) Hình ảnh đối tượng trong các ấn phẩm, bản in, bản vẽ.

vấn đối tượng trên hình ảnh cần phải trả về được các hình ảnh có chứa đối tượng quan tâm bát chấp mọi thay đổi trên bên ngoài của đối tượng. Những thay đổi đó có thể đến từ rất nhiều nguyên nhân khác nhau. Đó có thể do tác động từ các yếu tố bên ngoài khi chụp hình như điều kiện chiếu sáng, góc chụp của camera hay những tùy chỉnh khác nhau của các camera về độ tương phản, độ phân giải, màu sắc,... Cùng với đó là những hình ảnh của đối tượng được chụp với góc xoay, kích thước hình hay tỉ lệ khác nhau. Hoặc có những trường hợp đối tượng bị che khuất, cắt ghép, v.v... hoặc đối tượng được thể hiện trên các ấn phẩm, bản in, bản vẽ nên bị thay đổi về màu sắc và chi tiết. Một vài dạng thay đổi kể trên được

thể hiện qua Hình 1.2. Còn một trường hợp nữa là do những thay đổi từ chính bản thân đối tượng do các điều kiện bên ngoài ví dụ như đối tượng bị cũ đi hay bị xuống cấp theo thời gian.

Các loại đặc tính vật lý khác nhau trên mỗi đối tượng. Dựa trên các đặc tính vật lý người ta chia đối tượng thành các loại khác nhau. Có những đối tượng mà đặc tính thể hiện rõ nét nhất qua cấu trúc bề mặt, nhưng có cái lại qua màu sắc hay hình dạng, v.v... Ví dụ như với những con bướm, đặc trưng cho chúng không phải là hình dạng, kích cỡ vì đa phần các loài bướm đều có hình dạng, kích cỡ gần giống nhau mà ở đây là các họa tiết, màu sắc trên cánh bướm; Hay với những loại lá cây thì đặc trưng về màu sắc, họa tiết lại không cung cấp nhiều thông tin bằng hình dạng của lá.

Kích cỡ của tập dữ liệu lớn. Tập dữ liệu hình ảnh lớn thường bao gồm hàng triệu bức ảnh, vậy nên để người dùng có thể tương tác trực tiếp với hệ thống thông qua một thiết bị phía client như điện thoại di động thì đòi hỏi truy vấn phải được trả về trong thời gian ngắn chấp nhận được. Do đó cần phải có một thuật toán nhận dạng hiệu quả, chi phí thấp. Đồng thời những hình ảnh cũng phải được xử lý để lưu trữ sao cho tiết kiệm nhất để phù hợp với kích cỡ của RAM vì nếu lưu trữ trên ổ cứng sẽ mất rất nhiều thời gian để truy xuất và không thể đạt được yêu cầu về thời gian.

1.3 Mục tiêu và phạm vi của đề tài

Mục tiêu của luận văn này nhằm xây dựng một hệ thống truy vấn đối tượng trên ảnh từ tập dữ liệu lớn, trong đó quá trình truy vấn hoàn toàn dựa trên nội dung của ảnh và kết quả phải được trả về gần như ngay lập tức với cơ sở dữ liệu gồm hàng triệu hình ảnh chưa được gán nhãn. Hệ thống này tập trung vào giải quyết vấn đề về tìm kiếm một đối tượng cụ thể như một địa điểm, một bức tranh, một bìa sách, v.v... Những đối tượng này có thể được chụp trong các điều kiện khác nhau như góc chụp, ánh sáng, kích thước hay bị che khuất. Do đó mục đích của hệ thống không phải là trả về những hình ảnh chụp gần giống nhau như chụp trong cùng một khung cảnh hay cùng thuộc một loại đối tượng mà là trả về những hình ảnh có chứa chính xác đối tượng cần tìm. Ví dụ như khi đưa vào

một bức hình có chứa Nhà thờ Đức Bà, kết quả trả về sẽ những bức hình có chứa nhà thờ Đức Bà chứ không phải trả về những nhà thờ có kiến trúc hay có khung gian bao quanh giống với Nhà thờ Đức Bà.

1.4 Cấu trúc luận văn

Trong phần này, chúng tôi sẽ trình bày cấu trúc phần còn lại của luận văn và những vấn đề được thảo luận ở phần kế tiếp. Các nội dung sẽ được trình bày ở phần kế tiếp bao gồm:

Các công trình liên quan. Chúng tôi sẽ giới thiệu tổng quát về các công trình nghiên cứu liên quan tới truy vấn ảnh và bàn luận chi tiết về từng công trình trong Chương 2.

Các tập dữ liệu và phương pháp đánh giá. Để thử nghiệm kết quả của phương pháp đề xuất và so sánh hiệu suất của chúng với những phương pháp khác, chúng tôi thử nghiệm trên 3 bộ dữ liệu chuẩn là Oxford 5k, Paris 6k và Holiday. Kết quả sẽ được đánh giá bằng phương pháp mean Average Precision (mAP). Chi tiết của mỗi bộ dữ liệu cùng phương pháp đánh giá sẽ được trình bày chi tiết ở Chương 3.

Tích hợp thông tin không gian ảnh vào phương pháp đánh chỉ mục ngược. Chúng tôi đề xuất một phương pháp nhằm nâng cao hiệu suất của các hệ thống truy vấn đối tượng bằng cách tích hợp thông tin không gian ảnh vào phương pháp đánh chỉ mục ngược (inverted index). Trong Chương 4, chúng tôi sẽ trình bày chi tiết về ý tưởng của phương pháp, việc cài đặt cũng như kết quả thực nghiệm và đánh giá kết quả so với những phương pháp khác.

Tổng kết. Trong Chương 5, chúng tôi sẽ tổng kết, bàn luận thêm về phương pháp đề xuất và những đề xuất cải tiến, mở rộng để nâng cao hiệu suất của hệ thống trong thời gian tới.

Chương 2

Các công trình liên quan

Trong chương này chúng tôi sẽ trình bày một cách tổng quan về các phương pháp truy vấn đối tượng trên tập dữ liệu ảnh lớn đang được sử dụng rộng rãi hiện nay. Các phương pháp cần phải thỏa hai yêu cầu là cho kết quả với độ chính xác cao và trả về trong thời gian gần như ngay lập tức.

Để có thể truy vấn hình ảnh trong thời gian ngắn, mọi dữ liệu phải được lưu trữ trên RAM vì tốc độ truy xuất ổ cứng rất chậm. Tuy nhiên do dung lượng rất hạn chế của RAM, ta phải tìm cách biểu diễn tập dữ liệu hình ảnh cho phù hợp để vừa đảm bảo được về mặt không gian lưu trữ, vừa đáp ứng được các yêu cầu của truy vấn ảnh. Mục 2.1 sẽ trình bày ngắn gọn về hướng tiếp cận biểu diễn hình ảnh bằng các đặc trưng cục bộ. Nhưng khi kích cỡ của tập dữ liệu tăng thì việc so khớp các đặc trưng cục bộ tỏ ra kém hiệu quả. Trong mục 2.2, chúng tôi sẽ giới thiệu mô hình Bag-of-visual-words - được bắt nguồn từ mô hình Bag-of-Words (BoW) trong truy vấn văn bản. Mô hình này cho thấy tính hiệu quả của nó cả về tốc độ tính toán lẫn bộ nhớ sử dụng.

Mặc dù đạt được hiệu suất cao nhưng mô hình BoW vẫn bỏ qua thông tin về không gian ảnh - một thông tin quan trọng ảnh hưởng lớn đến độ chính xác của truy vấn. Trong mục 2.3, chúng tôi sẽ trình bày rõ hơn về các hướng tiếp cận dựa để khai thác được thông tin không gian ảnh, tiêu biểu là hướng tiếp cận dựa trên đặc trưng hình học và thông tin không gian của các đặc trưng cục bộ.

2.1 Biểu diễn hình ảnh bằng các đặc trưng cục bộ

Trong lĩnh vực Thị giác Máy tính, một câu hỏi và cũng là một thách thức lớn đối với tất cả các nhà khoa học là làm sao biểu diễn được một hình ảnh trên máy tính. Tùy theo từng mục đích cụ thể, người ta sẽ có các cách biểu diễn khác nhau. Trong truy vấn ảnh, một hình ảnh phải được biểu diễn dưới dạng sao cho bền vững trước những thay đổi như điều kiện chụp, tỉ lệ, góc chụp khác nhau hay thậm chí là những thay đổi lớn do đối tượng bị che khuất. Do sự tác động của các yếu tố này, cho dù hai hình ảnh chứa cùng một đối tượng thì vẫn có thể tồn tại một vùng hình ảnh lớn bên ngoài các đối tượng không đồng thời xuất hiện ở cả hai hình.

Để giải quyết vấn đề này, có một hướng tiếp cận phổ biến là rút trích những "chi tiết" cục bộ (local patches) trên tấm hình để biểu diễn cho hình ảnh đó. Hướng tiếp cận này được đưa ra dựa trên nhận định rằng hai hình ảnh tương tự nhau sẽ có rất nhiều những chi tiết cục bộ giống nhau và những chi tiết cục bộ này có thể được dùng để so khớp các hình ảnh với nhau. Các chi tiết này thường được rút trích bằng một trong hai phương pháp, đó là: (i) sử dụng một lưới dày đặc với nhiều mức tỉ lệ kích cỡ khác nhau (để đảm bảo bắt biết về tỉ lệ) để chia hình ảnh thành nhiều chi tiết nhỏ, hoặc (ii) dùng các phương pháp dò tìm (detector) hay một kỹ thuật nào đó để lấy được các chi tiết đặc biệt (đặc trưng) trên vùng hình ảnh quan tâm và đồng thời loại bỏ những chi tiết không đảm bảo sự bắt biến tỉ lệ ngay ở bước này. Có thể thấy rằng phương pháp dùng lưới để chia hình ảnh thành nhiều phần không thể áp dụng cho bài toán truy vấn ảnh với tập dữ liệu lớn vì ta cần rất nhiều không gian để lưu trữ một lượng lớn các chi tiết dày đặc với nhiều mức tỉ lệ kích cỡ khác nhau. Do vậy phương pháp biểu diễn hình ảnh bằng các đặc trưng được áp dụng cho bài toán này.

Có rất nhiều phương pháp dò tìm các đặc trưng (feature detector) được đưa ra, trong đó phải kể tới các phương pháp được dùng phổ biến như Difference of Gaussians, DoG ([Lowe \[2004\]](#)), Maximally Stable Extremal Regions, MSER ([Matas et al. \[2004\]](#)) và affine invariant detector ([Mikolajczyk & Schmid \[2004\]](#)). Ngoài ra còn có các phương pháp dò tìm được xây dựng để tìm kiếm trong thời

gian thực như SURF ([Bay et al. \[2006\]](#)), FAST ([Rosten et al. \[2010\]](#)) và BRISK ([Leutenegger et al. \[2011\]](#)).

Sau khi rút trích được các đặc trưng cục bộ cho mỗi hình, dựa trên các đặc trưng đó ta sẽ quyết định xem liệu hai tấm hình bất kỳ có chứa cùng một đối tượng hay không. Để so sánh độ tương đồng của hai đặc trưng cục bộ, ta không thể dựa trên màu sắc và cường độ của chúng vì những yếu tố này không bền vững trước những thay đổi của hình ảnh. Do đó ta cần phải tìm cách lượng tử hóa độ tương đồng giữa cách đặc trưng để có thể đo được bằng các tính toán cụ thể. Trong công trình nghiên cứu nổi tiếng của [Lowe \[2004\]](#), tác giả đã đề xuất một phương pháp để có thể tính toán được một bộ mô tả (descriptor) có tính phân loại cao và đảm bảo sự bất biến trước những thay đổi của hình ảnh, đó là SIFT descriptor. Theo sau công trình nghiên cứu này, nhiều công trình có hướng tiếp cận tương tự được đưa ra, trong đó bao gồm GLOH ([Mikolajczyk & Schmid \[2005\]](#)), SURF ([Bay et al. \[2006\]](#)), DAISY ([Tola et al. \[2008\]](#)), CONGAS ([Zheng et al. \[2009\]](#)), BRIEF ([Calonder et al. \[2010\]](#)). Đặc biệt, bằng việc đề xuất thuật toán RootSIFT được cải tiến từ SIFT, [Arandjelovic & Zisserman \[2012\]](#) đã nâng hiệu suất của phương pháp SIFT lên đáng kể. Đây cũng là phương pháp được chúng tôi chọn dùng trong hệ thống của mình.

Tóm lại, từ những bộ mô tả (descriptor) được rút trích từ tất cả các hình trong cơ sở dữ liệu và từ hình ảnh truy vấn, ta có thể tính toán được độ tương đồng giữa các hình ảnh. Tuy nhiên, hiệu suất của quá trình tính toán độ tương đồng bị giảm đi đáng kể khi thực hiện trên tập dữ liệu lớn. Trong phần tiếp theo, chúng tôi sẽ giới thiệu sơ lược về một mô hình giúp giải quyết được vấn đề này.

2.2 Mô hình Bag-of-words

Mô hình bag-of-words đã thể hiện được sức mạnh của nó trong truy vấn văn bản và được sử dụng trong các công cụ tìm kiếm văn bản mạnh mẽ như Google, Bing. Chính vì sự thành công đó, bag-of-words đã được sử dụng trong truy vấn ảnh. Mục này chủ yếu trình bày về việc ứng dụng phương pháp truy vấn văn bản này vào trong truy vấn ảnh. Trước tiên, chúng tôi sẽ sơ lược về truy vấn văn bản, tiếp đến sẽ là việc ứng dụng của nó trong truy vấn ảnh.

2.2.1 Truy vấn văn bản

Tương tự như hình ảnh, để có thể thực hiện truy vấn với văn bản, văn bản được biểu diễn dưới dạng một mô hình không gian vector ([Salton & McGill \[1986\]](#)) hay còn được gọi là mô hình *túi từ* (bag-of-words), BoW ([Manning et al. \[2008\]](#)). Theo đó, mỗi văn bản được xem như là một tập hỗn độn (một túi) các từ và được biểu diễn dưới dạng một biểu đồ (histogram) N_w -chiều với N_w là số các từ của một ngôn ngữ. Vì giá trị của mỗi cột cột của biểu đồ bằng với số lần xuất hiện của từ tương ứng với cột đó trong văn bản nên phương pháp này còn được gọi là *trọng số tần suất từ* (term frequency weighting).

Đôi khi, chúng ta có thể bắt gặp trường hợp nhiều từ xuất hiện trong các văn bản nhiều hơn các từ khác (ví dụ như trong tiếng Anh là *the* và *and*). Tuy nhiên những từ này thường mang ít giá trị hơn những từ ít phổ biến trong việc phục vụ cho mục đích so khớp. Do sự mất cân đối trong tần số xuất hiện của các từ, các chiều trong mô hình không gian vector phải được đánh trọng số dựa trên giá trị của thông tin mà từ đó mang chứ không phải dựa trên tần suất xuất hiện. Một phương pháp đánh trọng số thường được sử dụng là *tần số văn bản nghịch đảo*, idf (invert document frequency). Với N_D là tổng số các văn bản, N_i là số văn bản mà từ i xuất hiện, công thức tính *tần số văn bản nghịch đảo* được phát biểu như sau:

$$idf_i = \log \frac{N_D}{N_i} \quad (2.1)$$

Cuối cùng, trọng số của mỗi từ trong mỗi văn bản được tính bằng cách lấy tích của tần suất từ (term frequency - tf) và tần số nghịch đảo văn bản (invert document frequency - idf). Trọng số đó được gọi là tf-idf ([Manning et al. \[2008\]](#)) với công thức:

$$tf - idf_{i,d} = tf_{i,d} \times idf_i \quad (2.2)$$

Đối với những từ xuất hiện với tần suất cực kỳ lớn (stop word), ta có thể lọc và loại bỏ toàn bộ để giảm bớt chi phí về không gian lưu trữ và thời gian thực thi. Mức độ tương đồng giữa các văn bản sẽ được tính bằng công thức cosin áp dụng cho trọng số tf-idf của chúng trong mô hình bag-of-words. Thực tế mỗi văn bản

chỉ chứa một lượng rất nhỏ so với số lượng các từ có trong ngôn ngữ, do vậy vector sinh ra khi biểu diễn bằng mô hình bag-of-words sẽ rất thưa thớt. Để cho quá trình lưu trữ và truy vấn được hiệu quả, một cấu trúc dữ liệu sẽ được tính toán trước được gọi là *chỉ mục ngược* (inverted index) ([Manning et al. \[2008\]](#)). Chỉ mục ngược bao gồm một chuỗi các danh sách, mỗi danh sách tương ứng với một từ. Mỗi danh sách ghi lại những văn bản nào có chứa từ đó. Nhờ chỉ mục ngược, khi đưa vào một danh sách các từ rút từ văn bản truy vấn, ta có thể nhanh chóng lấy được danh sách các văn bản trong tập văn bản chứa các từ truy vấn đó. Từ đó có thể dễ dàng tính ra chỉ số tf-idf cho từng từ.

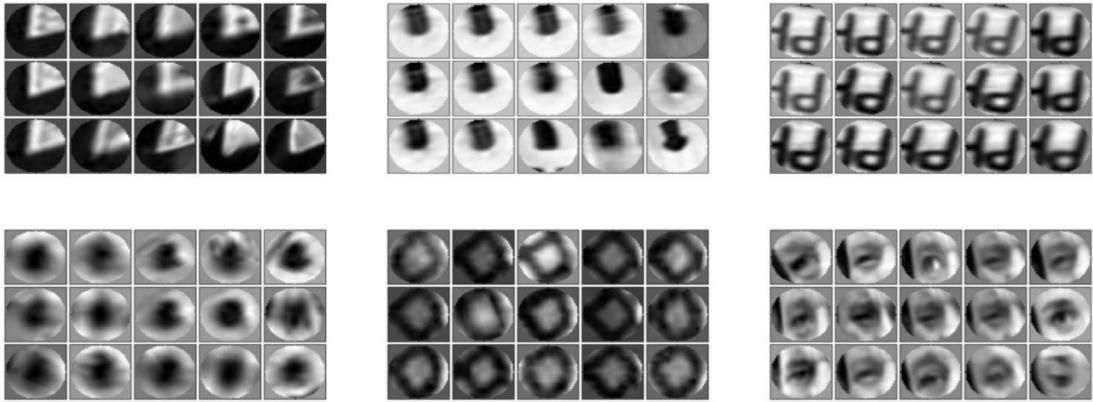
2.2.2 Bag-of-words trong truy vấn ảnh

Một khó khăn lớn khi áp dụng mô hình của truy vấn văn bản vào truy vấn ảnh là trong truy vấn văn bản, một văn bản có thể dễ dàng bóc tách ra các từ trong khi đó không có cách phân chia tự nhiên nào cho các hình ảnh. Như đã giới thiệu trong mục [2.1](#), một hình ảnh hoàn toàn có thể chia thành các đặc trưng cục bộ, tuy nhiên các đặc trưng này lại hoàn toàn phân biệt với nhau, vậy làm thế nào để xây dựng được các *từ* từ các đặc trưng này?

Nghiên cứu của [Sivic & Zisserman \[2003\]](#) là công trình đầu tiên ứng dụng hướng tiếp cận của truy vấn văn bản vào truy vấn ảnh¹. Trong công trình này tác giả đã giới thiệu khái niệm *các từ trực quan* (visual words) được tạo ra bằng cách sử dụng thuật toán gom cụm k-means để gom cụm các đặc trưng cục bộ. Hình [2.1](#) cho thấy một vài ví dụ về các từ trực quan. Tương tự như trong truy vấn văn bản, hình ảnh sẽ được rút trích các đặc trưng cục bộ rồi tiến hành gom cụm để biểu diễn thành các từ trực quan, sau đó được đánh trọng số bằng tf-idf, rồi biểu diễn dưới dạng mô hình bag-of-words và sử dụng chỉ mục ngược để tăng hiệu suất cho quá trình truy vấn. Thí nghiệm được tiến hành trên 4000 ảnh (frame) được lấy từ video và rút trích được 10,000 từ trực quan từ những hình ảnh đó.

Có một điều dễ thấy là nếu một hình ảnh được biểu diễn bằng càng nhiều từ trực quan thì hình ảnh đó càng "chi tiết" và độ chính xác của việc so khớp sẽ

¹Mục đích của tác giả trong nghiên cứu này là truy vấn trên video nhưng ta hoàn toàn có thể chuyển sang bài toán truy vấn ảnh bằng cách rút trích các frame trong video theo từng giây



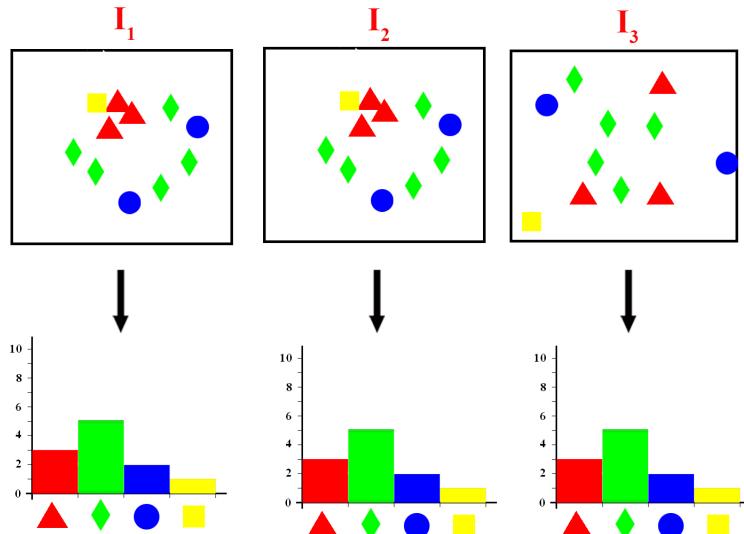
Hình 2.1: **Các từ trực quan (visual words).** Mỗi nhóm là một nhóm các đặc trưng cục bộ được rút trích từ hình ảnh, gom vào cùng một cụm và cùng được biểu diễn bằng một từ trực quan. Hình ảnh được lấy từ bài báo [Sivic & Zisserman, 2009].

tăng lên, đồng thời nó cũng khiến cho tốc độ truy vấn nhanh hơn vì các biểu đồ bag-of-words sẽ trở nên "thưa" hơn. Trong thực tế, để truy vấn ảnh trên những tập dữ liệu lớn, để cho kết quả tốt thì số lượng từ trực quan không thể vào khoảng 10,000 từ như trong thí nghiệm của Sivic & Zisserman [2003] mà phải lên tới hàng triệu từ. Trong khi đó, độ phức tạp của thuật toán k-means là $O(N_w N_d)$ với N_w , N_d lần lượt là số lượng từ trực quan và số lượng "văn bản" (hình ảnh) chứa chúng. Trên những tập dữ liệu lớn thì $N_d \geq N_w$ nên độ phức tạp luôn lớn hơn $O(N_w^2)$. Do đó ta không thể dùng k-means cho bài toán này. Nister & Stewenius [2006] đã đề xuất phương pháp giải quyết cho bài toán này bằng cách xây dựng một cây từ vựng mà về bản chất thì nó chính là thuật toán HKM (Hierarchical K-Means). Để minh họa cho thuật toán này, tác giả đã cho thử nghiệm trên bộ ảnh gồm 1 triệu hình ảnh. Không lâu sau đó, Philbin *et al.* [2007] đã đề xuất một hướng tiếp cận khác dựa trên thuật toán *xấp xỉ k-means*, AKM (Approximate K-Means). Tác giả cũng cho chạy thử nghiệm AKM trên 16.7 triệu đặc trưng để gom cụm thành 1 triệu từ. Các thí nghiệm cho thấy rằng, khi so sánh AKM với k-means thì về độ chính xác thì AKM xấp xỉ k-means tuy nhiên chi phí tính toán chỉ bằng một phần nhỏ của k-means. Còn khi so sánh AKM với HKM thì AKM không nhũng vượt xa về độ chính xác mà còn có thể áp dụng cho những tập dữ

liệu lớn. Chi phí tính toán của cả HKM và AKM đều là ($N_d \log(N_w)$).

2.3 Sử dụng thông tin không gian ảnh trong truy vấn ảnh

Mặc dù đạt được những kết quả rất đáng chú ý nhưng mô hình cơ bản của bag-of-words vẫn bị giới hạn về độ chính xác do bỏ qua một thông tin quan trọng, đó là thông tin về không gian của các đặc trưng cục bộ. Cấu trúc của mô hình bag-of-words như một cái túi chứa các từ một cách hỗn độn, không theo trật tự nên vị trí của các đặc trưng cục bộ xuất hiện trên hình không được chú ý đến, do đó các đặc trưng cục bộ được xử lý một cách rời rạc, không liên quan tới nhau. Hình 2.2 minh họa cho việc giảm độ chính xác của mô hình bag-of-words khi không chú ý tới thông tin không gian của các từ trực quan (visual words).



Hình 2.2: **Bỏ qua thông tin không gian ảnh trong mô hình bag-of-words.** Nếu bỏ qua thông tin không gian của các từ trực quan, ba hình ảnh trên sẽ được biểu diễn dưới dạng biểu đồ giống nhau do đó chúng sẽ được xem như ba hình ảnh giống nhau. Trong khi đó hình ảnh I_3 hoàn toàn khác với I_1 và I_2 .

Để giải quyết vấn đề trên, rất nhiều công trình nghiên cứu đã được đưa ra. Phần lớn các công trình nghiên cứu được chia ra làm hai dạng là tiếp cận dựa trên đặc trưng hình học và tiếp cận dựa trên thông tin không gian của các đặc trưng cục bộ. Mục 2.3.1 chúng tôi sẽ trình bày về các phương pháp dựa trên đặc trưng hình học. Còn hướng tiếp cận còn lại sẽ được trình bày chi tiết ở mục 2.3.2.

2.3.1 Các hướng tiếp cận dựa trên đặc trưng hình học

Các phương pháp sử dụng đặc trưng hình học để so khớp thường được dùng ở bước hậu xử lý để nhận dạng hình học. Dưới đây là một vài công trình tiêu biểu sử dụng hướng tiếp cận này.

Sivic & Zisserman [2003] đã đo đặc sự nhất quán không gian cục bộ (local spatial consistency) trong các so khớp giữa hình ảnh truy vấn và từng hình ảnh trong cơ sở dữ liệu từ đó tái xếp hạng lại danh sách kết quả trả về. Việc đo đặc sự nhất quán không gian cục bộ trong so khớp hình ảnh cũng được đề cập tới trước đó trong các công trình như Zhang *et al.* [1995] và Schmid *et al.* [1997].

Trong một công trình nghiên cứu của Philbin *et al.* [2007], tác giả sử dụng thuật toán RANSAC (Fischler & Bolles [1981]) để kiểm tra sự nhất quán hình học giữa các đặc trưng cục bộ trùng khớp. RANSAC là một trong những phương pháp phổ biến nhất cho hậu xử lý toàn cục trên hình ảnh. Đặc biệt, trong một công trình khác, Zhang *et al.* [2011] đề xuất mã hóa thông tin không gian ảnh qua các mệnh đề trực quan hình học (GVP) kết hợp với RANSAC đã cho kết quả rất đáng chú ý với bộ dữ liệu lên tới hàng triệu ảnh.

Trong khi đó, Lin & Brandt [2010] và Lampert [2009] lại xếp hạng các hình ảnh dựa trên điểm số so khớp của hình ảnh truy vấn với những cửa sổ con được định vị trên hình. Phương pháp này mã hóa được nhiều thông tin không gian ảnh hơn so với một hình bag-of-words trên toàn bộ tấm hình và giúp định vị hình ảnh truy vấn.

Nhìn chung, những phương pháp sử dụng hướng tiếp cận hình học đều cho kết quả tốt. Tuy nhiên, khi vùng truy vấn lớn hơn thì chúng chỉ được dùng để tái xếp hạng một số lượng giới hạn ở các hình ảnh ở top đầu của kết quả trả về vì vấn đề về chi phí cho bộ nhớ và tốc độ thực hiện.

2.3.2 Các hướng tiếp cận dựa trên thông tin không gian của các đặc trưng cục bộ

Hướng tiếp cận dựa trên đặc trưng hình học là hướng tiếp cận mang tính toàn cục, tức là xem xét đối tượng dưới một cái nhìn tổng quan, toàn thể chứ không xem xét chi tiết những thành phần cấu thành nó. Hướng tiếp cận dựa trên các đặc trưng cục bộ lại ngược lại, xem đối tượng là một tập hợp của nhiều thành phần và dựa trên những thành phần đó để xác định đối tượng. Lazebnik *et al.* [2006] đã giới thiệu một phương pháp nền tảng, được bắt nguồn từ ý tưởng *so khớp phân cấp* (pyramid matching) của Grauman & Darrell [2005], đó là phương pháp *so khớp không gian phân cấp* (Spatial Pyramid Matching - SPM). Ý tưởng của phương pháp này là lặp đi lặp lại việc chia nhỏ hình ảnh và tính toán biểu đồ của các đặc trưng cục bộ với mức độ chi tiết tăng dần. SPM đã giúp nâng cao một cách đáng kể độ chính xác cho mô hình bag-of-words và tỏ ra là một phương pháp đơn giản nhưng hiệu quả. Mặc dù vậy, SPM cũng làm tăng thời gian thực hiện truy vấn bởi khi mức độ chi tiết càng cao thì kích cỡ biểu đồ của các đặc trưng cục bộ cũng tăng theo làm tăng chi phí tính toán trong quá trình so khớp, vì vậy SPM vẫn chưa thích hợp cho các bài toán yêu cầu thời gian thực.

2.4 Kết chương

Việc biểu diễn hình ảnh bằng các đặc trưng cục bộ đã đặt nền tảng cho việc đưa ra các phương pháp để truy vấn đối tượng trên ảnh. Mô hình bag-of-words đã chứng minh tính hiệu quả của mình trong truy vấn ảnh và việc kết hợp phương pháp chỉ mục ngược (inverted index) giúp giảm đáng kể thời gian thực hiện truy vấn. Tuy nhiên, mô hình bag-of-words vẫn bị giới hạn về độ chính xác do bỏ qua thông tin không gian ảnh. Trong khi đó, rất nhiều hướng tiếp cận khác tận dụng được thông tin này đã nâng độ chính xác của truy vấn lên rất nhiều nhưng lại không quan tâm nhiều tới thời gian thực hiện.

Phương pháp chúng tôi đề xuất tập trung vào cả độ chính xác và thời gian truy vấn. Để đạt được mục đích đó, chi phí bộ nhớ cao có thể được chấp nhận.

Chương 3

Phương pháp đề xuất

Bài toán mà luận văn này tập trung giải quyết là truy vấn đối tượng trên tập dữ liệu lớn trong thời gian gần với thời gian thực. Rất nhiều công trình được đưa ra để giải quyết bài toán này (mục 2.1 và 2.2). Phương pháp cơ bản để giải quyết bài toán là biểu diễn một hình ảnh dưới dạng mô hình bag-of-visual-words (BoW), sau đó xếp hạng các hình ảnh sử dụng phương pháp tf-idf và dùng chỉ mục ngược (inverted index) để tăng hiệu suất tính toán. Tuy nhiên, phương pháp trên vẫn còn bị giới hạn về độ chính xác do chưa sử dụng đến thông tin không gian ảnh. Các phương pháp được đưa ra trong những năm gần đây để giải quyết vấn đề này đã được giới thiệu trong mục 2.3.

Trong chương này, chúng tôi sẽ mô tả phương pháp đề xuất để tích hợp thông tin không gian ảnh vào chỉ mục ngược. Trước tiên chúng tôi sẽ nhắc lại những công trình khơi nguồn ý tưởng cho phương pháp của chúng tôi. Sau đó là phần trình bày chi tiết phương pháp đề xuất.

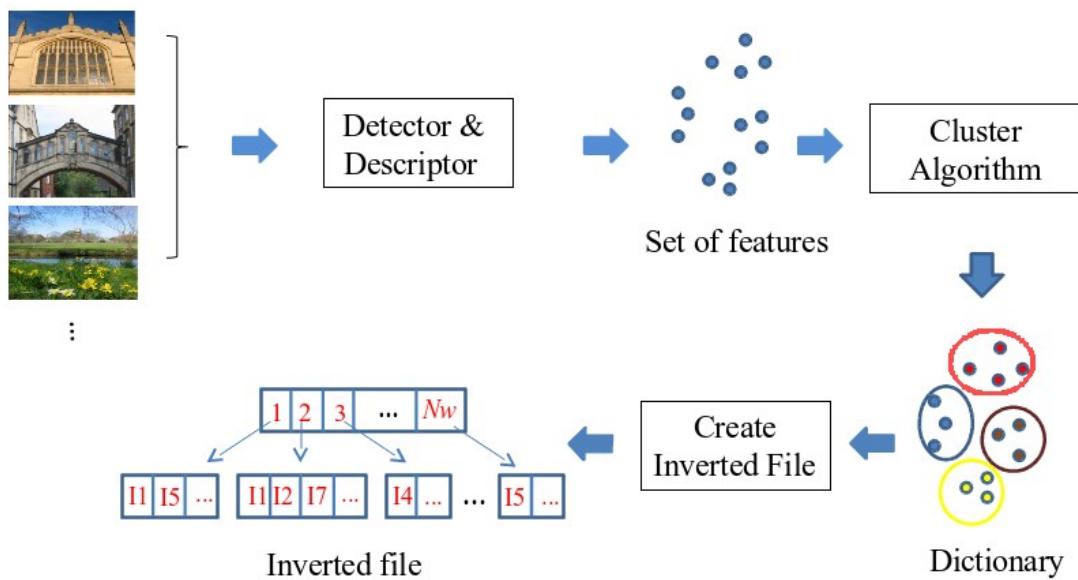
3.1 Chỉ mục ngược với biểu diễn Bag-of-Visual-Words

Như đã được giới thiệu sơ lược trong mục 2.2.1, chỉ mục ngược (inverted index) là phương pháp phổ dùng để tối ưu hóa tốc độ truy vấn cơ sở dữ liệu bằng việc lưu trữ trước một ánh xạ từ nội dung đến vị trí trong cơ sở dữ liệu. Nói cách

3. Tích hợp thông tin không gian ảnh vào chỉ mục ngược

khác, chỉ mục ngược là một cấu trúc dữ liệu chủ yếu bao gồm 2 trường là khóa và giá trị. Mỗi khóa đại diện cho một từ, và phần giá trị của tương ứng lưu trữ danh sách các văn bản có chứa từ đó. Vì vậy ta có thể dễ dàng lấy được danh sách tất cả các văn bản chứa từ truy vấn.

Chính vì sự thành công của các kỹ thuật tìm kiếm văn bản, chỉ mục ngược đã được mở rộng để sử dụng cho tìm kiếm ảnh trên cơ sở dữ liệu lớn. Để có thể xây dựng chỉ mục ngược cho cơ sở dữ liệu ảnh, mô hình BoW đã được sử dụng để biểu diễn hình ảnh. Quá trình xây dựng chỉ mục ngược như sau: (i) một bộ dò tìm các đặc trưng sẽ phát hiện những điểm quan trọng, sau đó một bộ mô tả sẽ trích rút trích được những đặc trưng xung quanh điểm đó; (ii) các đặc trưng được gom thành các cụm để tạo thành từ điển, mỗi cụm là một tập các đặc trưng gần giống nhau và trung tâm của mỗi cụm là một *từ trực quan* (visual word), mỗi từ trực quan sẽ được gán một mã số khác nhau; (iii) Trường giá trị trong tệp chỉ mục ngược sẽ lưu trữ danh sách các hình ảnh có chứa các từ trực quan tương ứng. Quá trình tạo tập tin chỉ mục ngược (inverted file) được minh họa trong hình 3.1.



Hình 3.1: Quá trình tạo tập tin chỉ mục ngược (inverted index file).

3. Tích hợp thông tin không gian ảnh vào chỉ mục ngược

Trong quá trình truy vấn, các đặc trưng sẽ được rút trích từ hình ảnh truy vấn, sau đó từ các đặc trưng ta sẽ lấy được các từ trực quan bằng cách sử dụng từ điển sau đó tra cứu trong tập tin chỉ mục ngược để lấy được các hình ảnh ứng viên. Những hình ảnh nào có số lượng từ trực quan trùng với các từ trong hình ảnh truy vấn càng nhiều thì sẽ càng được xếp hạng cao hơn trong danh sách kết quả truy vấn trả về. Kỹ thuật này được gọi là *bầu chọn* (voting).

Bên cạnh kỹ thuật bầu chọn, để nâng cao độ chính xác của kết quả trả về, người ta có thể thêm một bước tái xếp hạng danh sách kết quả bằng cách tính khoảng cách trong không gian đặc trưng giữa hình ảnh truy vấn và các hình ảnh ứng viên sử dụng biểu diễn BoW của chúng. Tuy nhiên, chi phí tính toán của quá trình này rất cao dẫn đến thời gian thực hiện truy vấn tăng đáng kể.

Trong thí nghiệm được trình bày ở Chương 4, chúng tôi sẽ so sánh cả hai phương pháp bầu chọn và tái xếp hạng với phương pháp được đề xuất.

3.2 Tích hợp thông tin không gian ảnh vào chỉ mục ngược

Phương pháp chúng tôi đề xuất nhằm tích hợp thông tin không gian ảnh vào chỉ mục ngược được bắt nguồn từ ý tưởng của một công trình nghiên cứu của [Lazebnik et al. \[2006\]](#). Trong công trình đó, thay vì sử dụng một biểu đồ (histogram) chung của các từ trực quan để biểu diễn một hình ảnh thì họ chia hình ảnh thành các nhiều phần sử dụng lưới ô vuông phân cấp (hay còn được gọi là không gian phân cấp - spatial pyramid). Một lưới ô vuông tại cấp l sẽ chia hình ảnh thành $2^l \times 2^l$ ô với kích cỡ như nhau. Do đó, số ô vuông trên lưới ở cấp 0 là 1×1 ; cấp 1 là 2×2 . Nếu cấp l càng cao thì lưới ô vuông sẽ càng dày đặc hơn. Nếu coi mỗi ô của hình ảnh được chia bởi lưới ô vuông phân cấp là một hình ảnh độc lập, dựa trên mô hình BoW ta sẽ tính được các biểu đồ độc lập. Chính vì mức độ chia tiết của các biểu đồ khác nhau nên chúng sẽ được đánh trọng số khác rồi rồi được ghép nối với nhau để tạo thành một vector đặc trưng biểu diễn cho hình ảnh. Bằng cách biểu diễn như vậy, các hình ảnh có sự phân bố các từ tương tự nhau

3. Tích hợp thông tin không gian ảnh vào chỉ mục ngược

sẽ được biểu diễn bằng những biểu đồ ghép nối gần giống nhau.

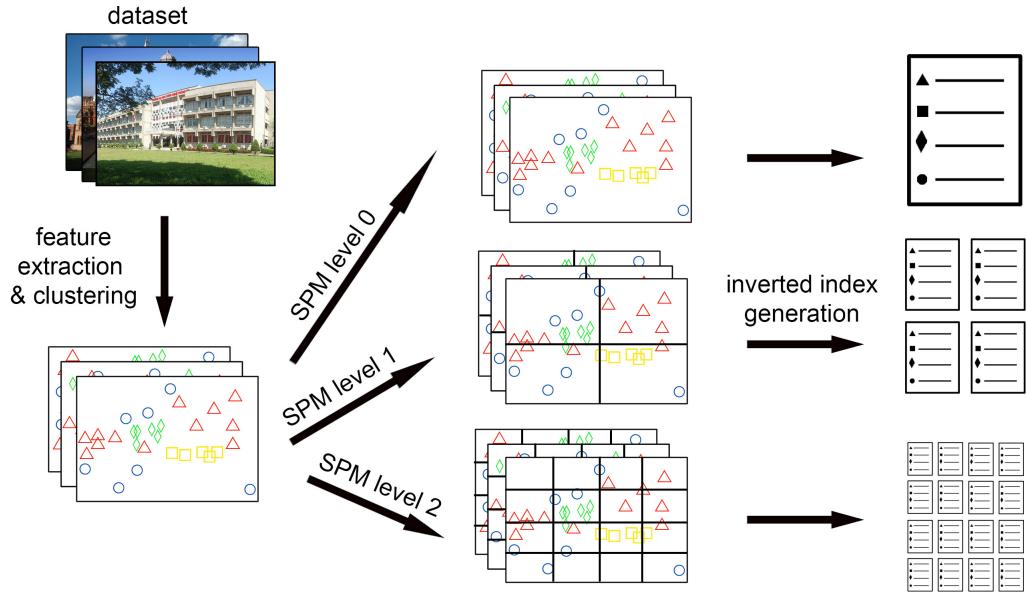
Dựa trên ý tưởng của nghiên cứu trên, chúng tôi đề xuất sử dụng không gian phân cấp để tăng cường mức độ bầu chọn và lập chỉ mục của kỹ thuật đánh chỉ mục ngược căn bản. Ý tưởng được chúng tôi đưa ra là chia hình ảnh thành nhiều ô sử dụng không gian phân cấp và giới hạn ở một cấp xác định. Sau đó các từ trực quan sẽ được đánh số tương ứng với các ô chúng rơi vào. Ta sẽ duyệt qua tất cả các ô ở tất cả các cấp khác nhau để thực hiện việc bầu chọn. Do đó, nếu hai hình ảnh chứa các từ trực quan giống nhau trong cùng một ô sẽ nhận được nhiều lượt bầu chọn hơn so với hai hình ảnh có các từ trực quan giống nhau nhưng lại nằm rải rác ở các ô khác nhau. Các lượt bầu chọn sẽ được đánh trọng số tùy theo từng cấp. Nếu cấp càng cao hay diện tích của mỗi ô càng hẹp thì trọng số của lượt bầu chọn càng cao. Trọng số tại cấp l sẽ là $\frac{1}{2^{L-l}}$.

Một trong những điểm đặc biệt của phương pháp đề xuất là chúng tôi sử dụng đa chỉ mục ngược. Tức là chia thành nhiều tập tin chỉ mục ngược khác nhau nhưng các tập tin vẫn giữ được cấu trúc căn bản của chỉ mục ngược. Mỗi tập tin sẽ dùng để lưu trữ chỉ mục cho một ô trên không gian phân cấp. Nếu cấp độ cao nhất của không gian phân cấp là L thì tổng số lượng tập tin chỉ mục ngược sẽ là $\frac{1}{3}(4^{L+1} - 1)$ và mỗi cấp độ sẽ có $2^l \times 2^l$ tập tin chỉ mục ngược với $0 \leq l \leq L$. Hình 3.2 mô tả khái quát cho phương pháp được đề xuất.

Khi thực hiện quá trình rút trích các đặc trưng cho tất cả các hình trong cơ sở dữ liệu, thông tin không gian của các đặc trưng đó sẽ được lưu trữ lại. Sau đó các bộ mô tả (descriptors) của đặc trưng (ví dụ như key points) sẽ được lượng tử hóa để tạo thành một bảng từ vựng của các từ trực quan (từ điển). Mỗi hình ảnh sẽ chứa một tập các từ trực quan. Tiếp đó ta sẽ sử dụng không gian phân cấp để chia tất cả các hình ảnh thành các ô nhỏ với "độ mịn" tăng dần dựa trên cấp được định nghĩa. Lúc này, thông tin không gian của các đặc trưng đã được lưu trữ trước đó sẽ được sử dụng để xác định xem từ đó có thuộc ô đang xét hay không. Tất cả các từ được tìm thấy trong mỗi ô sẽ được thu thập lại. Tiếp theo, tập hợp của các từ được tìm thấy trong mỗi ô của các hình ảnh sẽ được dùng để sinh ra một tập tin chỉ mục ngược tương ứng với ô đó. Số lượng tập tin chỉ mục ngược được sinh ra bằng với tổng số ô của không gian phân cấp.

Trong quá trình truy vấn, các đặc trưng cũng được rút trích từ hình ảnh truy vấn. Sau đó chúng được đưa vào từ điển để lấy được các từ trực quan tương ứng.

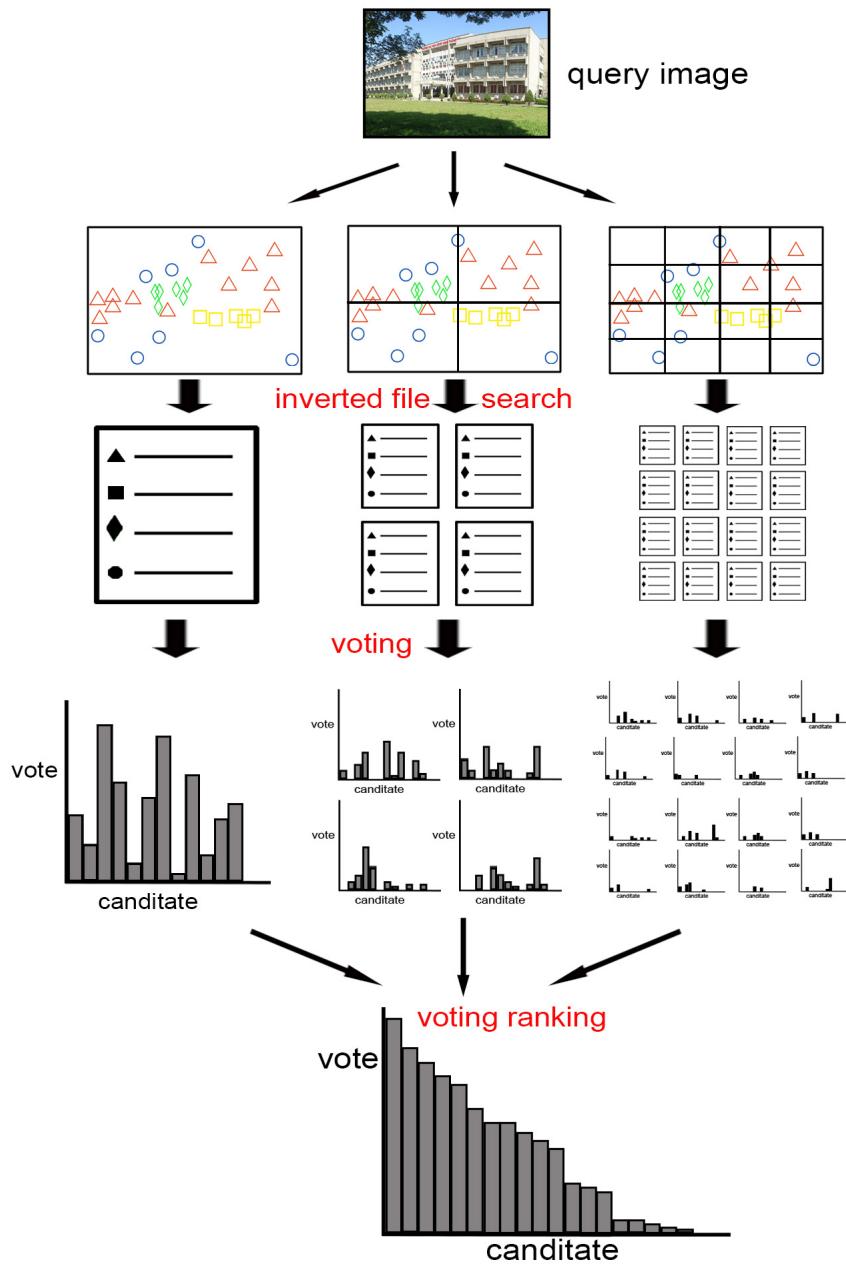
3. Tích hợp thông tin không gian ảnh vào chỉ mục ngược



Hình 3.2: Khái quát về phương pháp đề xuất.

Dựa vào vị trí của các từ này, ta có thể xác định được chúng thuộc ô nào tại mỗi cấp của mô hình không gian phân cấp. Từ đó ta có thể có thể truy xuất ngay lập tức tới tập tin chỉ mục ngược tương ứng với mỗi ô để lấy và xếp hạng danh sách hình ảnh ứng viên một cách đồng thời. Ta xếp hạng hình ảnh bằng phương pháp bầu chọn nên việc bầu chọn diễn ra trong mỗi lần truy xuất tập tin chỉ mục ngược, do đó danh sách đếm số lượt bầu chọn sẽ được cập nhật liên tục trong suốt quá trình truy xuất các tập tin chỉ mục ngược. Khi quá trình bầu chọn kết thúc, ta sẽ tổng hợp toàn bộ số lượt bầu chọn cho từng hình rồi xếp hạng các hình theo số lượt bầu chọn. Toàn bộ quá trình truy vấn của phương pháp đề xuất được minh họa trong Hình 3.3.

3. Tích hợp thông tin không gian ảnh vào chỉ mục ngược



Hình 3.3: Quá trình truy vấn của phương pháp đề xuất.

Chương 4

Thực nghiệm và đánh giá kết quả

Chương này sẽ trình bày...

4.1 Các bộ dữ liệu và phương thức đánh giá

Mục này sẽ trình bày quy trình đánh giá chuẩn được sử dụng rộng rãi trong truy vấn đối tượng trên tập dữ liệu lớn. Trước tiên là mô tả về các bộ dữ liệu chuẩn, sau đó là phần trình bày chi tiết về phương thức đánh giá cho các kết quả thí nghiệm.

4.1.1 Các bộ dữ liệu

4.1.1.1 Oxford 5k

Bộ dữ liệu Oxford 5k được xây dựng bởi Philbin *et al.* [2007], bao gồm 11 Oxford "landmark"¹ cùng các hình ảnh gây nhiễu. Hình ảnh cho mỗi landmark được tự động lấy về từ trang chia sẻ ảnh trực tuyến Flickr sử dụng các câu truy vấn như "Oxford Christ Church" và "Oxford Radcliffe Camera", đồng thời các hình ảnh gây nhiễu cũng được lấy về bằng câu truy vấn "Oxford". Bộ dữ liệu bao gồm 5,062 hình ảnh chất lượng cao (1366×768).

Tập dữ liệu đánh giá chuẩn (ground truth) được xây dựng thủ công cho 11 landmark. Các hình ảnh được gán vào một trong bốn nhãn: *Good* nếu nó là một

¹Landmark ở đây có nghĩa là một góc nhìn/góc chụp đặc biệt của một tòa nhà

4. Các bộ dữ liệu và phương thức đánh giá

hình ảnh rõ ràng và đầy đủ về đối tượng/tòa nhà, *OK* nếu hình ảnh chứa hơn 25% của đối tượng và *Junk* nếu hình ảnh chứa ít hơn 25% của đối tượng hoặc đối tượng bị che khuất phần lớn hoặc hình ảnh đối tượng bị méo mó nhiều.

Bộ dữ liệu gồm 55 truy vấn trong đó mỗi landmark sẽ có 5 truy vấn. Các đối tượng sẽ được khoanh vùng trên các hình ảnh truy vấn. Tất cả các truy vấn được thể hiện trong hình 4.1.

4.1.1.2 Paris 6k

Tương tự như bộ dữ liệu Oxford 5k, Paris 6k bao gồm 6,392 hình ảnh chất lượng cao (1366×768) của các địa danh nổi tiếng ở Paris được lấy về từ Flickr với các câu truy vấn như "Paris Eiffel Tower" hay "Paris Triomphe". Paris 6k cũng có 55 hình ảnh truy vấn cho 11 landmark (5 truy vấn cho mỗi landmark) ([Philbin et al. \[2008\]](#)).

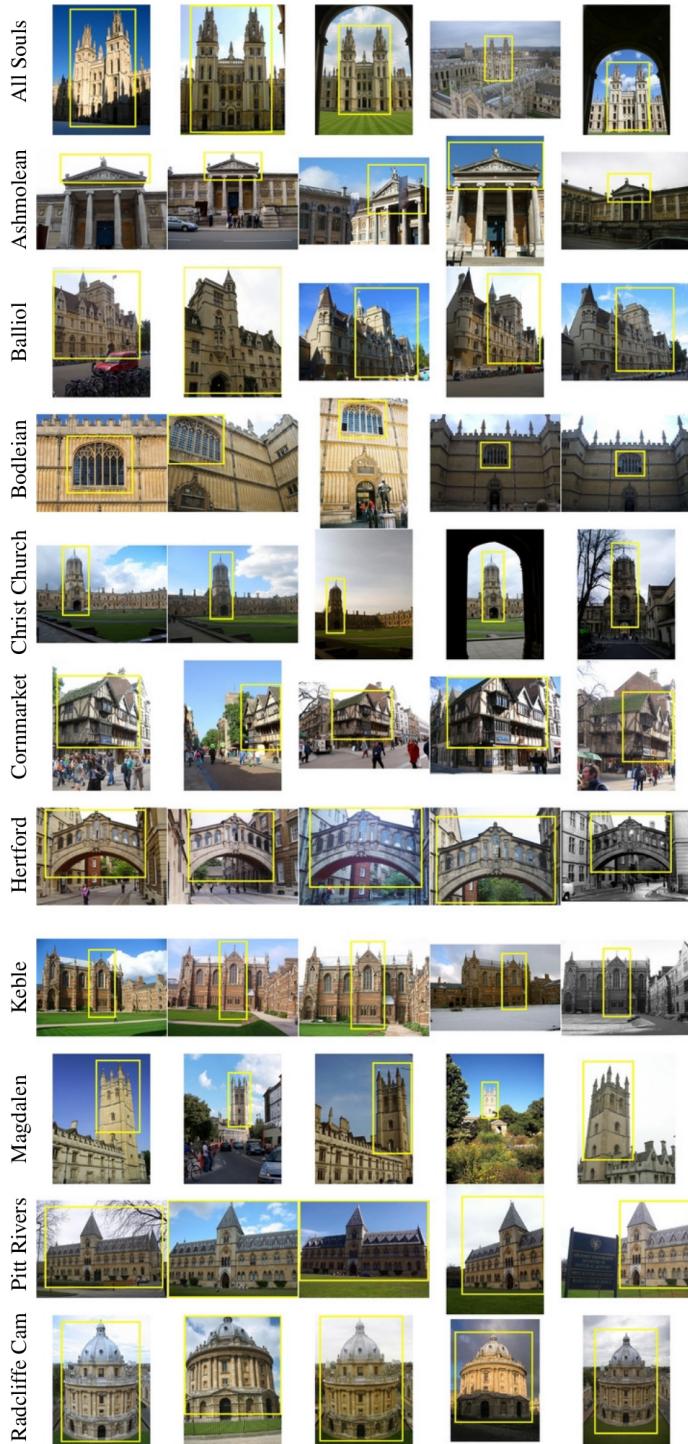
Paris 6k được đánh giá là một bộ dữ liệu hoàn toàn độc lập so với Oxford 5k và thường được dùng để kiểm tra các tác động của việc tính toán từ trực quan trong khi Oxford 5k thường được dùng để kiểm tra hiệu suất.

4.1.1.3 Holidays

Bộ Holidays bao gồm 1,491 hình ảnh chất lượng cao về các lễ hội với 500 truy vấn mà mỗi truy vấn chứa một vài hình ảnh chính xác về đối tượng truy vấn ([Jégou et al. \[2008\]](#)). Bộ dữ liệu này nhằm hướng tới mục đích truy vấn ảnh trên tập dữ liệu lớn vì các hình ảnh không phải về một đối tượng đặc biệt như trong Oxford 5k và Paris 6k, đồng thời các thay đổi trên hình ảnh của đối tượng cũng đa dạng hơn với nhiều loại thay đổi khác nhau và truy vấn được thực hiện với toàn bộ hình ảnh chứ không phải với một vùng hình ảnh có chứa đối tượng được chỉ định trước.

Tập dữ liệu bao gồm 500 nhóm hình ảnh, mỗi nhóm là một cảnh khác nhau và bao gồm rất nhiều loại hình ảnh khác nhau như tự nhiên, nhân tạo, hiệu ứng nước và lửa,... Hình ảnh đầu tiên của mỗi nhóm là hình truy vấn và các hình còn lại trong nhóm là kết quả chính xác cho truy vấn. Và các hình ảnh truy vấn không được xem xét tới trong kết quả trả về chứ không phải được xem là một

4. Các bộ dữ liệu và phương thức đánh giá



Hình 4.1: **Landmark và các truy vấn** được dùng để đánh giá 55 hình ảnh truy vấn được sử dụng trong tập dữ liệu đánh giá chuẩn. Mỗi hàng là 5 hình của 5 truy vấn khác nhau cho cùng một cảnh landmark. Hình ảnh được lấy từ bài báo [Philbin *et al.*, 2007].

4. Các bộ dữ liệu và phương thức đánh giá

kết quả chính xác như trong Oxford 5k và Paris 6k.

Thông thường bộ dữ liệu Holidays thường được ghép với khoảng 1 triệu hình ảnh từ Flickr khác để kiểm tra truy vấn trên tập dữ liệu lớn, thường được gọi là bộ dữ liệu *Holidays + Flickr1M*.

4.1.2 Phương thức đánh giá

Với mỗi truy vấn, để đánh giá kết quả trả về ta thường dùng độ đo *precision-recall* (PR). Precision là tỉ lệ giữa số kết quả đúng trả về trong tổng số kết quả trả về. Recall là tỉ số của số kết quả đúng trả về trên tổng số hình ảnh đúng trong tập dữ liệu. Hay nói theo cách khác, precision cho thấy độ "tinh khiết" của kết quả trả về, còn recall cho biết đã tìm thấy bao nhiêu phần của đáp án.

Tùy theo từng mục đích mà người ta sẽ tập trung vào việc nâng cao precision hay recall. Ví dụ, những ứng dụng như Google Goggles¹ thì câu hỏi nó cần phải trả lời là "Nó là cái gì?", do đó nó chỉ chú ý đến việc đạt được chỉ số precision tối đa có thể, tức là lấy được những kết quả đúng nhưng vừa đủ để nhận dạng đối tượng. Trong nhiều trường hợp khác thì chỉ số recall cũng được quan tâm. Ví dụ việc tái tạo không gian ba chiều đòi hỏi phải tìm được đủ số lượng hình ảnh của đối tượng để xây dựng được mô hình ba chiều chính xác.

Để đo hiệu suất thực thi của hệ thống, ở đây ta dùng độ đo average precision (AP)(Philbin *et al.* [2007]), nó cũng tương đương với phần diện tích bên dưới đường biểu diễn cho chỉ số precision-recall trong biểu đồ. Một đường biểu diễn precision-recall lý tưởng có chỉ số precision bằng 1 trên tất cả các mức recall khác nhau và nó tương ứng chỉ số average precision bằng 1. Average precision được tính cho từng truy vấn một sau đó ta lấy trung bình cộng của chúng, đó chính là mean Average Precision - một con số để đánh giá hiệu suất tổng thể của hệ thống.

¹Google Goggles là một ứng dụng nhận dạng hình ảnh được phát hành bởi Google. Người sử dụng điện thoại di động chỉ cần chụp ảnh của đối tượng như xe hơi, đồ chơi, bìa sách, mã vạch,... sau đó Goggles sẽ quét và đối chiếu kho dữ liệu để hiển thị thông tin liên quan đến vật đó.

4. Các bộ dữ liệu và phương thức đánh giá

4.2 Cài đặt thí nghiệm

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum euismod libero a augue suscipit, a dignissim justo ullamcorper. Nunc in tincidunt dolor, sed feugiat ante. Curabitur ut elit sit amet dui euismod congue porttitor rhoncus enim. Donec in mollis massa, et sagittis est. Duis est dui, suscipit id sollicitudin vel, malesuada non metus. Vivamus tincidunt libero non nunc dignissim, interdum auctor lorem lacinia. Praesent ultrices nec turpis placerat consectetur. Maecenas volutpat lobortis interdum. Fusce ullamcorper nunc at varius bibendum. Praesent at ipsum sagittis, facilisis leo ac, commodo orci.

Proin eu velit semper, molestie ipsum sit amet, consequat quam. Fusce rhoncus est in facilisis mattis. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vivamus accumsan, erat sed egestas luctus, orci turpis posuere dolor, ut tempor magna enim et mauris. Maecenas in scelerisque quam, id bibendum lacus. Nullam rutrum odio id magna porttitor tincidunt. Cum sociis natoque penatibus et magnis dis parturient montes, nascentur ridiculus mus. Aliquam feugiat elit sapien, eget consectetur sapien tempor et. Aenean tempus eleifend laoreet. Vivamus nec mollis orci. Maecenas sit amet quam nibh. Phasellus blandit fringilla massa faucibus porta. Fusce mattis pellentesque leo, sed auctor mi lobortis viverra. Donec sed arcu non orci gravida accumsan nec in diam. Proin malesuada enim sed est sagittis, tincidunt sollicitudin metus suscipit. Vivamus dapibus suscipit diam, fringilla fermentum arcu dignissim eu.

Aliquam eget velit vitae ligula pharetra malesuada. Maecenas ut facilisis lorem, in dapibus dolor. Curabitur pulvinar dolor a adipiscing dignissim. Maecenas sed porttitor ligula. Morbi vitae lacus laoreet, posuere odio eget, vehicula dui. Pellentesque lectus metus, rutrum vulputate nisl et, consequat rutrum augue. Fusce mollis dolor vitae lectus commodo ornare. Sed volutpat in magna ut mollis. Maecenas sodales tincidunt iaculis. Interdum et malesuada fames ac ante ipsum primis in faucibus. Interdum et malesuada fames ac ante ipsum primis in faucibus. Proin ut aliquam lorem. Suspendisse adipiscing lacinia dictum. Morbi at augue id mauris imperdiet tincidunt eu sit amet elit.

4. Các bộ dữ liệu và phương thức đánh giá

4.3 Kết quả thí nghiệm và đánh giá kết quả

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean tincidunt risus eros, ac molestie quam lobortis at. Proin ante dolor, lacinia vel risus cursus, commodo rutrum lacus. Donec dapibus euismod sollicitudin. Sed viverra sapien tempor velit pulvinar, a condimentum quam aliquam. Vivamus purus purus, sagittis eu bibendum in, vulputate nec magna. Aliquam erat volutpat. Nulla facilisi.

Sed porta elit in vehicula pellentesque. Morbi faucibus mollis libero, ac volutpat lectus sagittis ut. Vestibulum eget fermentum eros. Suspendisse potenti. Nunc ac luctus nunc, id dapibus dolor. Curabitur lorem ante, pretium et nunc nec, iaculis laoreet metus. Pellentesque vitae nisi id magna pulvinar pulvinar. Quisque venenatis dolor sit amet velit elementum tincidunt. Sed purus dui, varius in dui eget, mollis venenatis lacus. Fusce vestibulum metus eget mauris accumsan varius. Donec dapibus iaculis cursus. Duis ut congue diam. Quisque convallis mi sodales, condimentum nisi nec, consequat lectus. Integer posuere venenatis hendrerit.

Chương 5

Kết luận và hướng phát triển

5.1 Kết luận

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus id quam at nisi imperdiet ultrices. Suspendisse porttitor velit non neque suscipit, eget aliquet ante pharetra. Pellentesque in ornare orci, vitae adipiscing eros. Proin eu purus sollicitudin, imperdiet augue vel, elementum arcu. Cras condimentum, sapien et vestibulum cursus, turpis risus sollicitudin mauris, quis ultricies nibh nibh sed lorem. Nulla ante dolor, sodales in libero et, cursus auctor nisi. Mauris felis tortor, bibendum sed hendrerit eget, rhoncus vitae magna. In hac habitasse platea dictumst. Suspendisse fringilla viverra nibh, ut viverra odio venenatis vel. Fusce ante mauris, laoreet non elementum ac, gravida a neque. In at mauris urna. Quisque vitae consectetur est. Mauris neque est, varius id interdum quis, lobortis imperdiet tortor. Quisque sapien massa, facilisis nec sapien quis, suscipit ultrices mi. Vestibulum ut felis ac sapien scelerisque tristique.

In ullamcorper magna massa, ac mollis ligula viverra quis. Mauris velit dui, luctus in massa sit amet, tincidunt faucibus felis. Duis cursus fermentum tortor, in facilisis velit tincidunt a. Nullam felis orci, faucibus vel ultrices at, vestibulum non tellus. Donec eu adipiscing mi. Mauris semper turpis diam, vel consectetur metus tincidunt a. Donec euismod elementum enim. Nulla quam mi, lacinia quis gravida eget, imperdiet et libero.

5.2 Hướng phát triển

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam lacus nisl, ullamcorper a condimentum quis, eleifend in tortor. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Ut nec congue felis. Cras convallis nulla a dui aliquet mollis. Sed eget convallis enim. Morbi sed aliquet ante. Integer ut nisi a ipsum sodales condimentum et et dolor. Aliquam et commodo massa. Praesent lobortis, nunc non fermentum iaculis, mauris elit cursus velit, non fringilla purus magna nec nunc. Praesent tincidunt venenatis nibh nec tempus. Mauris mattis tortor nunc, quis fermentum velit mollis in. Curabitur eu nibh enim. Nam sodales vulputate quam eget lacinia. Duis ultrices erat eu molestie auctor. Aliquam erat volutpat.

Phụ lục A

Đây là phụ lục A

Phụ lục B

Đây là phụ lục B

References

- ARANDJELOVIC, R. & ZISSEMAN, A. (2012). Three things everyone should know to improve object retrieval. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2911–2918, IEEE. [9](#)
- BAY, H., TUYTELAARS, T. & VAN GOOL, L. (2006). Surf: Speeded up robust features. In *Computer Vision-ECCV 2006*, 404–417, Springer. [9](#)
- CALONDER, M., LEPESTIT, V., STRECHA, C. & FUA, P. (2010). Brief: Binary robust independent elementary features. In *Computer Vision-ECCV 2010*, 778–792, Springer. [9](#)
- FISCHLER, M.A. & BOLLES, R.C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, **24**, 381–395. [14](#)
- GRAUMAN, K. & DARRELL, T. (2005). The pyramid match kernel: Discriminative classification with sets of image features. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2, 1458–1465, IEEE. [15](#)
- JÉGOU, H., DOUZE, M. & SCHMID, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In A.Z. David Forsyth Philip Torr, ed., *European Conference on Computer Vision*, vol. I of *LNCS*, 304–317, Springer. [23](#)
- LAMPERT, C.H. (2009). Detecting objects in large image collections and videos by efficient subimage retrieval. In *Computer Vision, 2009 IEEE 12th International Conference on*, 987–994, IEEE. [14](#)

REFERENCES

- LAZEBNIK, S., SCHMID, C. & PONCE, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2169–2178, IEEE. [15](#), [18](#)
- LEUTENECKER, S., CHLI, M. & SIEGWART, R.Y. (2011). Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2548–2555, IEEE. [9](#)
- LIN, Z. & BRANDT, J. (2010). A local bag-of-features model for large-scale object retrieval. In *Computer Vision–ECCV 2010*, 294–308, Springer. [14](#)
- LOWE, D.G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, **60**, 91–110. [8](#), [9](#)
- MANNING, C.D., RAGHAVAN, P. & SCHÜTZE, H. (2008). *Introduction to information retrieval*, vol. 1. Cambridge university press Cambridge. [10](#), [11](#)
- MATAS, J., CHUM, O., URBAN, M. & PAJDLA, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, **22**, 761–767. [8](#)
- MIKOŁAJCZYK, K. & SCHMID, C. (2004). Scale & affine invariant interest point detectors. *International journal of computer vision*, **60**, 63–86. [8](#)
- MIKOŁAJCZYK, K. & SCHMID, C. (2005). A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **27**, 1615–1630. [9](#)
- NISTER, D. & STEWENIUS, H. (2006). Scalable recognition with a vocabulary tree. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2, 2161–2168, IEEE. [12](#)
- PHILBIN, J., CHUM, O., ISARD, M., SIVIC, J. & ZISSERMAN, A. (2007). Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 1–8, IEEE. [12](#), [14](#), [22](#), [24](#), [25](#)

REFERENCES

- PHILBIN, J., CHUM, O., ISARD, M., SIVIC, J. & ZISSERMAN, A. (2008). Lost in quantization: Improving particular object retrieval in large scale image databases. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8, IEEE. [23](#)
- ROSTEN, E., PORTER, R. & DRUMMOND, T. (2010). Faster and better: A machine learning approach to corner detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **32**, 105–119. [9](#)
- SALTON, G. & MCGILL, M.J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA. [10](#)
- SCHMID, C., MOHR, R. et al. (1997). Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19**, 530–534. [14](#)
- SIVIC, J. & ZISSERMAN, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 1470–1477, IEEE. [11](#), [12](#), [14](#)
- SIVIC, J. & ZISSERMAN, A. (2009). Efficient visual search of videos cast as text retrieval. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **31**, 591–606. [12](#)
- TOLA, E., LEPETIT, V. & FUA, P. (2008). A fast local descriptor for dense matching. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 1–8, IEEE. [9](#)
- ZHANG, Y., JIA, Z. & CHEN, T. (2011). Image retrieval with geometry-preserving visual phrases. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 809–816, IEEE. [14](#)
- ZHANG, Z., DERICHE, R., FAUGERAS, O. & LUONG, Q.T. (1995). A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial intelligence*, **78**, 87–119. [14](#)

REFERENCES

- ZHENG, Y.T., ZHAO, M., SONG, Y., ADAM, H., BUDDEMEIER, U., BISSACCO, A., BRUCHER, F., CHUA, T.S. & NEVEN, H. (2009). Tour the world: building a web-scale landmark recognition engine. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1085–1092, IEEE. 9