

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

**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**

**CỬ NHÂN NGÀNH KHOA HỌC MÁY TÍNH**

**GIẢNG VIÊN HƯỚNG DẪN:**

**TS. NGÔ ĐỨC THÀNH  
PGS. TS. LÊ ĐÌNH DUY**

**TP. HỒ CHÍ MINH, 2014**

## LỜI CẢM ƠN

Để hoàn thành luận văn này, trước tiên chúng em xin gửi lời cảm ơn chân thành nhất đến hai người thầy đã dùu dắt, tận tình hướng dẫn chúng em là thầy Ngô Đức Thành và thầy Lê Đình Duy. Các thầy đã truyền cho chúng em không chỉ là kiến thức mà còn là niềm đam mê, trách nhiệm và sự nỗ lực hết mình trong công việc để có thể hoàn thành tốt luận văn này. Chúng em hi vọng sẽ còn tiếp tục được học hỏi nhiều thêm nữa từ các thầy trong tương lai.

Tiếp theo chúng em xin cảm ơn các thầy cô, bạn bè thuộc Phòng thí nghiệm Truyền thông Đa phương tiện, trường Đại học Công nghệ Thông tin - ĐHQG Tp. HCM, những người đã giúp đỡ, góp ý và đồng hành với chúng em trong suốt quá trình hoàn thành khóa luận này. Những bài học chúng em học được trong suốt quãng thời gian làm việc tại phòng thí nghiệm sẽ là hành trang quý báu cho chúng em khi ra trường.

Và không thể thiếu được trong sự thành công của chúng con là nguồn động lực, tình yêu thương từ phía gia đình, ba mẹ, những người đã sinh thành và nuôi dưỡng chúng con lớn khôn như ngày hôm nay.

Mặc dù đã cố gắng hết sức mình để hoàn thiện luận văn một cách tốt nhất có thể nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Kính mong nhận được sự góp ý và chỉ bảo tận tình từ phía thầy cô và các bạn.

Nhóm thực hiện  
Nguyễn Văn Biên & Phạm Duy

## 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 kho dữ liệu ảnh có sẵn. 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, bảo vệ thương hiệu, tìm kiếm sản phẩm, v.v... Tuy nhiên, bài toán trên cũng đang đối mặt với nhiều thách thức. Bên cạnh sự đa dạng về biểu hiện hình ảnh của cùng một 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 ngắn khi tìm kiếm trong tập dữ liệu lớn.

Rất nhiều công trình nghiên cứu đã được đề xuất để giải quyết vấn đề trên. Hầu hết các công trình này đều dựa trên mô hình Bag-of-Words (BoW), theo đó mỗi hình ảnh sẽ được biểu diễn bằng các điểm đặc trưng (interest points), sau đó các điểm đặc trưng này được lượng tử hóa vào các visual word. Cuối cùng mỗi hình ảnh sẽ được biểu diễn bằng một vector đặc trưng, là histogram của 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 về phân bố trong không gian ảnh của các điểm đặc trưng.

Trong luận văn này, chúng tôi đề xuất một phương pháp nhằm tích

hợp thông tin trên 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 Oxford 100K 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

<b>Mục lục</b>	<b>iv</b>
<b>Danh sách hình vẽ</b>	<b>vii</b>
<b>Danh sách bảng</b>	<b>ix</b>
<b>Danh sách từ viết tắt</b>	<b>x</b>
<b>1 Tổng quan</b>	<b>1</b>
1.1 Đặt vấn đề . . . . .	1
1.2 Thách thức . . . . .	3
1.3 Mục tiêu, đối tượng và phạm vi nghiên cứu . . . . .	5
1.3.1 Mục tiêu . . . . .	5
1.3.2 Đối tượng nghiên cứu . . . . .	5
1.3.3 Phạm vi nghiên cứu . . . . .	6
1.4 Cấu trúc luận văn . . . . .	6
<b>2 Các công trình liên quan</b>	<b>8</b>
2.1 Mô hình tổng quan . . . . .	8
2.2 Rút trích đặc trưng hình ảnh . . . . .	9
2.2.1 Đặc trưng toàn cục . . . . .	10
2.2.1.1 Đặc trưng hình dạng . . . . .	10
2.2.1.2 Đặc trưng texture . . . . .	12
2.2.1.3 Đặc trưng màu sắc . . . . .	13
2.2.2 Đặc trưng cục bộ . . . . .	14
2.3 Biểu diễn hình ảnh bằng mô hình Bag-of-words . . . . .	16

## MỤC LỤC

---

2.4	So khớp hình ảnh . . . . .	18
2.5	Sử dụng thông tin về sự phân bố trong không gian ảnh của các đặc trưng . . . . .	21
2.5.1	Các hướng tiếp cận dựa trên đặc trưng hình học . . . . .	22
2.5.2	Các hướng tiếp cận dựa trên thông tin không gian của các điểm đặc trưng cục bộ . . . . .	23
2.6	Kết chương . . . . .	24
<b>3</b>	<b>Phương pháp đề xuất</b>	<b>25</b>
3.1	Chỉ mục ngược trong truy vấn hình ảnh . . . . .	25
3.2	Tích hợp thông tin không gian ảnh vào chỉ mục ngược . . . . .	27
<b>4</b>	<b>Thực nghiệm và đánh giá kết quả</b>	<b>31</b>
4.1	Các bộ dữ liệu và phương thức đánh giá . . . . .	31
4.1.1	Các bộ dữ liệu . . . . .	32
4.1.1.1	Oxford 5K . . . . .	32
4.1.1.2	Paris 6K . . . . .	32
4.1.1.3	Oxford 5K+100K . . . . .	32
4.1.2	Phương thức đánh giá . . . . .	34
4.2	Cài đặt thí nghiệm . . . . .	35
4.2.1	Các phương pháp đánh giá cùng thông số cài đặt . . . . .	35
4.2.2	Nâng cao hiệu suất của hệ thống . . . . .	36
4.2.2.1	Tăng độ chính xác của quá trình gom cụm . . . . .	37
4.2.2.2	Lọc bỏ các stop word . . . . .	38
4.3	Kết quả thí nghiệm và đánh giá kết quả . . . . .	38
<b>5</b>	<b>Xây dựng ứng dụng</b>	
	<b>tìm kiếm đối tượng trên ảnh</b>	<b>46</b>
5.1	Tổng quan ứng dụng . . . . .	46
5.1.1	Mục đích và phạm vi của ứng dụng . . . . .	46
5.1.2	Các chức năng chính . . . . .	47
5.2	Xây dựng ứng dụng . . . . .	47
5.2.1	Kiến trúc tổng quan . . . . .	47
5.2.2	Server side . . . . .	49

## **MỤC LỤC**

---

5.2.2.1	Quá trình training . . . . .	49
5.2.2.2	Quá trình truy vấn . . . . .	50
5.2.3	Web service . . . . .	51
5.2.4	Client side . . . . .	54
5.2.4.1	Tổng quan . . . . .	54
5.2.4.2	Chức năng và giao diện . . . . .	57
5.2.5	Tối ưu hiệu suất hệ thống . . . . .	57
5.3	Cài đặt và thực nghiệm . . . . .	63
5.3.1	Môi trường cài đặt . . . . .	63
5.3.2	Kết quả thực nghiệm và đánh giá ứng dụng . . . . .	63
<b>6</b>	<b>Kết luận và hướng phát triển</b>	<b>66</b>
6.1	Kết luận . . . . .	66
6.2	Hướng phát triển . . . . .	67
6.2.1	Mở rộng phương pháp đề xuất . . . . .	67
6.2.2	Phát triển ứng dụng thực tế . . . . .	67
<b>Tài liệu tham khảo</b>		<b>69</b>

# Danh sách hình vẽ

1.1	Ví dụ minh họa về bài toán tìm kiếm ảnh . . . . .	2
1.2	Những thay đổi bè ngoài của đối tượng trên ảnh . . . . .	4
2.1	Mô hình tổng quan của một hệ thống truy vấn ảnh . . . . .	9
2.2	Biểu diễn hình ảnh bằng mô hình Bag-of-words . . . . .	16
2.3	Mô hình Bag-of-words . . . . .	17
2.4	Các từ visual word . . . . .	18
2.5	Bỏ qua thông tin về sự phân bố trong không gian của các visual word trong mô hình BoW . . . . .	21
2.6	Phương pháp Spatial Pyramid Matching . . . . .	23
3.1	Quá trình tạo chỉ mục ngược . . . . .	26
3.2	Khái quát về phương pháp đề xuất . . . . .	29
3.3	Quá trình truy vấn của phương pháp đề xuất . . . . .	30
4.1	Landmark và các truy vấn được dùng để đánh giá . . . . .	33
4.2	Thông số cài đặt thí nghiệm của mô hình BoW . . . . .	36
4.3	Biểu đồ so sánh độ chính xác và thời gian truy vấn giữa các phương pháp . . . . .	40
4.4	Biểu đồ so sánh độ chính xác và thời gian truy vấn giữa các phương pháp . . . . .	42
4.5	Biểu đồ hiệu suất của phương pháp đề xuất trên các cấp độ phân cấp khác nhau . . . . .	43
4.6	Ví dụ thể hiện kết quả khi tìm kiếm một hình ảnh trên bộ dữ liệu Oxford 5K . . . . .	44

## **DANH SÁCH HÌNH VẼ**

---

4.7 Ví dụ thể hiện kết quả khi tìm kiếm một hình ảnh trên bộ dữ liệu Paris 6K . . . . .	45
5.1 Kiến trúc tổng quan của hệ thống . . . . .	48
5.2 Quá trình truy vấn tại server . . . . .	51
5.3 Sơ đồ mô hình hoạt động chi tiết của web service . . . . .	53
5.4 Kiến trúc của client side . . . . .	54
5.5 Sơ đồ tổ chức các lớp của ứng dụng trên hệ điều hành Android. . . . .	55
5.6 Hình ảnh ứng dụng với các chức năng chính . . . . .	58
5.7 Hình ảnh ứng dụng trong khi thực hiện truy vấn . . . . .	58
5.8 Hình ảnh ứng dụng khi có kết quả trả về . . . . .	59
5.9 Hình ảnh ứng dụng khi xem chi tiết kết quả truy vấn . . . . .	60
5.10 Hình ảnh ứng dụng với chức năng tải thêm hình ảnh trong danh sách kết quả . . . . .	61
5.11 Hình ảnh ứng dụng với chức năng thay đổi chế độ xem ảnh . . . . .	61
5.12 Kết quả thực nghiệm hiệu suất của hệ thống tìm kiếm đối tượng trên ảnh. . . . .	64

# Danh sách bảng

4.1	Số hình ảnh trong mỗi bộ dữ liệu và số truy vấn trong tập dữ liệu đánh giá chuẩn tương ứng . . . . .	34
4.2	So sánh kết quả truy vấn với số lần lặp khác nhau trong thuật toán gom cụm AKM . . . . .	37
4.3	Kết quả lọc bỏ các stop words . . . . .	38
4.4	Hiệu suất của các phương pháp trên bộ dữ liệu Oxford 5K . . . . .	39
4.5	Hiệu suất của các phương pháp trên bộ dữ liệu Paris 6K . . . . .	40
4.6	Hiệu suất của các phương pháp trên bộ dữ liệu Holidays . . . . .	41
4.7	Hiệu suất của phương pháp đề xuất với các giá trị khác nhau của $L$ trên bộ dữ liệu Oxford 5K . . . . .	42
4.8	Hiệu suất của phương pháp đề xuất với các giá trị khác nhau của $L$ trên bộ dữ liệu Paris 6K . . . . .	43

# Danh mục từ viết tắt

**BoW** Bag-of-Words

**SPM** Spatial Pyramid Matching

**AP** Average Precision

**mAP** mean Average Precision

**DoG** Difference of Gaussians

**tf-idf** term frequency-inverse document frequency

**HKM** Hierarchical K-Means

**AKM** Approximate K-Means

# Chương 1

## Tổng quan

### 1.1 Đặt vấn đề

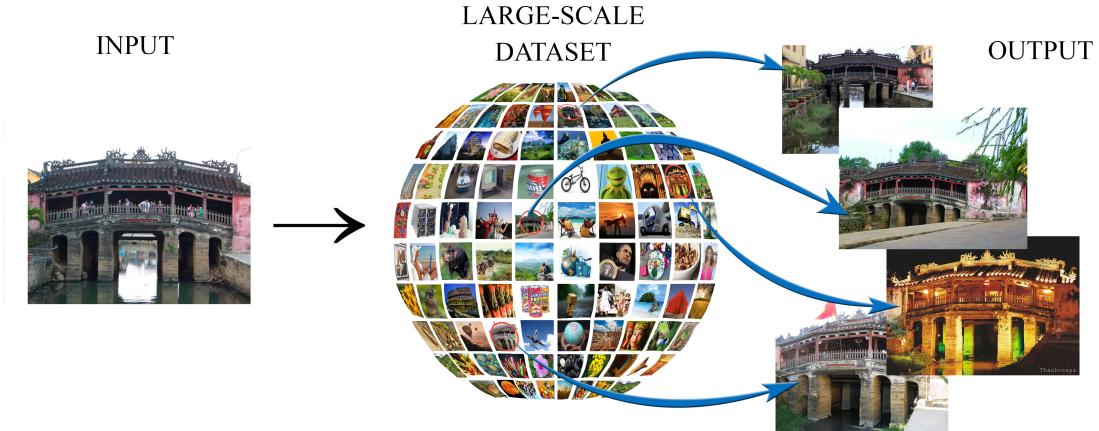
Trong những năm gần đây, cùng với sự phát triển của công nghệ thông tin, các lĩnh vực liên quan đến kỹ thuật số cũng đang có tốc độ phát triển chóng mặt. Các thiết bị kỹ thuật số như máy ảnh, máy quay phim kỹ thuật số, camera số, điện thoại di động có chức năng chụp hình,... đang ngày càng phổ biến và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

Dể đáp ứng yêu cầu đó, rất nhiều hệ thống truy vấn ảnh đã ra đời. Với đầu vào là một hình ảnh có chứa đối tượng quan tâm, hệ thống sẽ trả về những hình ảnh có chứa đối tượng quan tâm trong kho dữ liệu ảnh có sẵn. 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. Chúng tôi sẽ liệt kê sơ lược một vài ứng dụng của hệ thống này dưới đây.

#### \* 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



Hình 1.1: Ví dụ minh họa về bài toán tìm kiếm ảnh

dụng cụ thẻ:

**Tìm kiếm 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...

**Xác định 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 hay ta muốn biết một đối tượng xuất hiện trong cảnh nào của một bộ phim, v.v... Bởi vì video được tạo nên từ các khung hình khác nhau nên một hệ thống truy vấn ảnh sẽ dễ dàng thực hiện điều này chỉ với một hình ảnh của đối tượng quan tâm. **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 tìm kiếm thông tin về 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.

**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 vấn đề thách thức phải giải quyết. Dưới đây chúng tôi sẽ trình bày những thách thức chính trong bài toán này:

**Sự biến đổi bề ngoài của đối tượng trong hình ảnh.** Một hệ thống truy vấn đối tượng trên hình ảnh cần phải trả về được các hình ảnh có chứa đối tượng quan tâm bất 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,



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. (Những hình ảnh này được lấy từ Google Image Search với query là "chùa Một Cột").

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.

**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 giải pháp tìm kiếm hiệu quả, chi phí thấp. Đồng thời những hình ảnh cũng phải được xử lý để lưu trữ sao cho tiết kiệm nhất cho phù hợp với nhiều thiết bị với tài nguyên tính toán khác nhau.

### 1.3 Mục tiêu, đối tượng và phạm vi nghiên cứu

#### 1.3.1 Mục tiêu

Mục tiêu của khóa luận 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. Hệ thống cần có độ chính xác tương đương hoặc tốt hơn so với các giải pháp phổ biến nhất hiện nay trong khi vẫn có thể duy trì tốc độ phản hồi cao. Hệ thống này tập trung vào giải quyết vấn đề về tìm kiếm một đối tượng cụ thể. 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 không gian bao quanh giống với Nhà thờ Đức Bà.

#### 1.3.2 Đối tượng nghiên cứu

Đối tượng nghiên cứu trong khóa luận này là các hệ thống truy vấn ảnh, trong đó tập trung vào các giải pháp cho rút trích các đặc trưng hình ảnh, các phương

pháp biểu diễn hình ảnh trên máy tính để phục vụ cho mục đích truy vấn, các kỹ thuật giúp tăng tốc quá trình truy vấn, các phương pháp sử dụng thông tin không gian ảnh trong bài toán truy vấn để nâng cao độ chính xác của truy vấn. Cùng với đó là các phương pháp và các bộ dữ liệu chuẩn được sử dụng rộng rãi trên thế giới để đánh giá kết quả của phương pháp đề xuất cho bài toán truy vấn.

### **1.3.3 Phạm vi nghiên cứu**

Đề tài tập trung nghiên cứu những vấn đề chính sau:

- Các kiến thức nền tảng trong lĩnh vực xử lý ảnh về phát hiện và rút trích các đặc trưng trên ảnh.
- Các kỹ thuật biểu diễn hình ảnh.
- Các kỹ thuật, phương pháp nâng cao trong bài toán truy vấn ảnh: kỹ thuật đánh chỉ mục, kỹ thuật so khớp hình ảnh, phương pháp xếp hạng hình ảnh.
- Các bộ dữ liệu chuẩn để đánh giá hiệu suất của các phương pháp cho bài toán tìm kiếm đối tượng trên ảnh.

## **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 quan về các công trình nghiên cứu có liên quan tới bài toán truy vấn ảnh, các hướng tiếp cận và bàn luận chi tiết về từng công trình trong [Chương 2](#).

**Phương pháp đề xuất.** 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 về phân bố trong không gian của các điểm đặc trưng vào phương pháp đánh chỉ mục ngược (inverted index). Chi tiết về phương pháp đề xuất được trình bày trong [Chương 3](#).

**Thí nghiệm.** Để đánh giá hiệu quả của phương pháp đề xuất và so sánh với các phương pháp khác, chúng tôi tiến hành thí nghiệm trên ba bộ dữ liệu chuẩn là Oxford 5K, Paris 6K và Oxford 100K. Kết quả được tính bằng phương pháp

## Chương 1. Tổng quan

mean Average Precision (mAP). Chương 4 sẽ trình bày chi tiết về các bộ dữ liệu, việc cài đặt thí nghiệm và kết quả thí nghiệm.

**Xây dựng ứng dụng thực nghiệm.** Chương 5 trình bày việc xây dựng, cài đặt và thực nghiệm hệ thống truy vấn đối tượng trên ảnh. Việc xây dựng ứng dụng nhằm chứng minh tính thực tế của đề tài.

**Tổng kết.** Trong Chương 6, chúng tôi sẽ tổng kết về luận văn, bàn luận thêm về phương pháp đề xuất và hướng 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.

Trước tiên, chúng tôi sẽ trình bày tổng quan về các thành phần của một hệ thống truy vấn ảnh cơ bản trong mục 2.1. Mục 2.2 trình bày về các phương pháp rút trích đặc trưng hình ảnh dựa trên hai hướng tiếp cận là đặc trưng toàn cục và đặc trưng cục bộ. Mục 2.3 giới thiệu một mô hình dựa trên hướng tiếp cận đặc trưng cục bộ, đó là mô hình Bag-of-Words (BoW). Quá trình so khớp hình ảnh được biểu diễn bằng mô hình BoW sẽ được trình bày tại mục 2.4. Cuối cùng, chúng tôi sẽ trình bày trong mục 2.5 về các hướng tiếp cận khai thác 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 Mô hình tổng quan

Mô hình tổng quan của một hệ thống truy vấn cơ bản ảnh gồm ba thành phần chính:

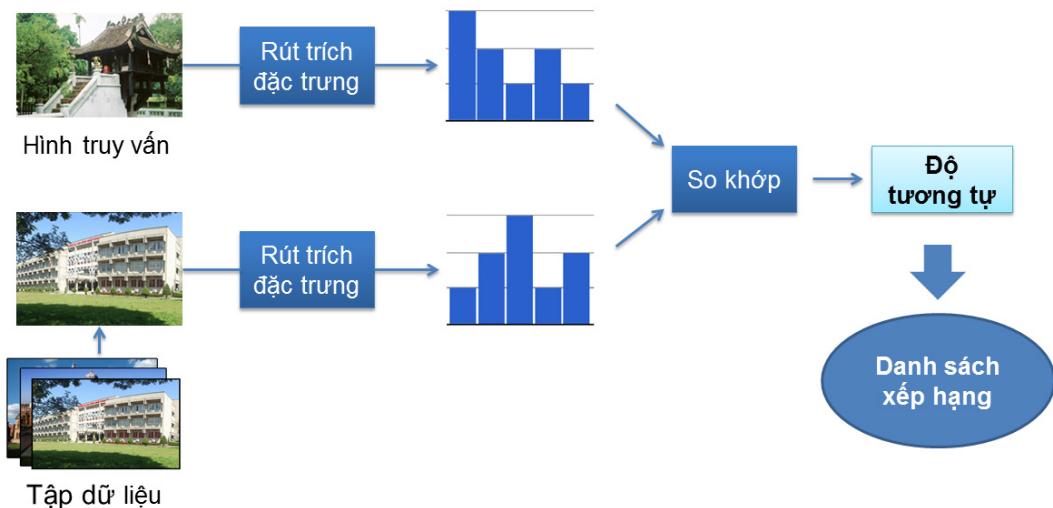
**Rút trích, biểu diễn đặc trưng ảnh.** Từ một hình ảnh, hệ thống sẽ dò tìm và phát hiện được những điểm đặc trưng, sau đó các điểm này sẽ được mô tả để rút ra được một vector tương ứng với mỗi điểm. Từ các vector đó, ta sẽ xây dựng được một histogram biểu diễn cho hình ảnh đó.

**So khớp các hình ảnh.** Sau khi biểu diễn các hình ảnh dưới dạng các histogram,

ta tiến hành so khớp các histogram của các hình trong cơ sở dữ liệu với của hình truy vấn để tìm được những hình ảnh có độ tương đồng cao nhất so với hình ảnh truy vấn. Các hình ảnh sẽ được xếp hạng dựa trên độ tương đồng này.

**Hậu xử lý kết quả.** Từ danh sách có thứ tự các hình ảnh có độ tương đồng cao nhất so với hình ảnh truy vấn, ta tiến hành xử lý tùy theo mục đích cụ thể của hệ thống hoặc có thể áp dụng thêm một thuật toán để re-ranking những hình ảnh trong top đầu nhằm tăng độ chính xác.

Mô hình tổng quan của một hệ thống truy vấn ảnh cơ bản được minh họa trong hình 2.1.



Hình 2.1: Mô hình tổng quan của một hệ thống truy vấn ảnh.

## 2.2 Rút trích đặc trưng hình ảnh

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. Điều này gây khó khăn cho việc so khớp các hình ảnh.

Trong phần này, chúng tôi sẽ trình bày về hai hướng tiếp cận phổ biến được dùng để giải quyết vấn đề. Đó là hướng tiếp cận dựa trên đặc trưng toàn cục và đặc trưng cục bộ.

### 2.2.1 Đặc trưng toàn cục

Để nhận biết các vật thể quanh ta bằng thị giác, ta thường dựa trên các đặc trưng dễ nhận thay bằng mắt thường như màu sắc, hình dạng hay texture bề ngoài của vật. Đó chính là những đặc trưng toàn cục gần gũi nhất với thị giác của con người. Trong mục này, chúng tôi sẽ lần lượt trình bày các công trình nghiên cứu về những đặc trưng toàn cục cơ bản sử dụng trong truy vấn ảnh đó là: đặc trưng hình dạng, đặc trưng về texture và đặc trưng màu sắc.

#### 2.2.1.1 Đặc trưng hình dạng

Hình dạng là một trong những đặc trưng quan trọng có thể nhìn thấy và cũng là một đặc trưng cơ bản để mô tả nội dung của hình ảnh. Tuy nhiên, biểu diễn và mô tả hình dạng của đối tượng trên ảnh là một việc vô cùng khó khăn bởi vì khi một đối tượng không gian ba chiều được chiếu lên mặt phẳng không gian hai chiều sẽ làm mất đi một chiều thông tin của đối tượng. Do đó, hình dạng của đối tượng rút trích được từ hình ảnh thường chỉ thể hiện được một phần hình dạng của đối tượng. Ngoài ra, còn một vài khó khăn nữa phải đối mặt trong việc rút trích đặc trưng về hình dạng như về độ nhiễu, sự mất mát thông tin hình ảnh, sự che khuất hay méo mó, v.v...

Các kỹ thuật biểu diễn và mô tả hình dạng thường được chia làm hai dạng:  
**Biểu diễn hình dạng dựa trên đường viền.** Kỹ thuật này chỉ tập trung khai thác thông tin về hình dạng đường biên của đối tượng trên ảnh. Nó được chia ra làm hai loại là tiếp cận liên tục và tiếp cận rời rạc.

- Tiếp cận liên tục: là không chia hình dạng thành nhiều phần nhỏ, thông thường một vector sẽ được rút ra từ toàn bộ đường bao dùng để mô tả

## Chương 2. Các công trình liên quan

---

hình dạng. Và độ tương đồng giữa các hình dạng sẽ được đo bằng cách tính khoảng cách giữa các vector này.

Các bộ mô tả đơn giản phổ biến bao gồm *diện tích*, *độ tròn* ( $C^2/S$ , trong đó C là chu vi, S là diện tích), *tính tâm sai* (chiều dài của trục lớn / chiều dài của trục nhỏ), *hướng của trục chính*, *lượng uốn cong* (bending energy)[1]. Những bộ mô tả đơn giản này thường chỉ có thể phân biệt hình dạng sai khác lớn, do đó chúng thường được dùng để lọc những trường hợp sai và kết hợp với những bộ mô tả hình dạng khác để phân biệt các hình dạng. Những bộ mô tả hình dạng đường viền khác được đề xuất bởi Peura and Iivarinen[2] bao gồm độ lỗi, tỉ lệ của các trục cơ bản, phương sai đường tròn và phương sai elip.

Một phương pháp khác cũng được dùng phổ biến là phương pháp so khớp dựa trên sự tương xứng hình dạng. Phương pháp này ngược với kỹ thuật biểu diễn hình dạng dựa trên đặc trưng, tức là nó đo độ tương tự giữa các hình dạng sử dụng so khớp từng điểm. Nói cách khác, mọi điểm trên hình dạng được coi như các điểm đặc trưng. Việc so khớp được thực hiện trên không gian 2D. Một trong những phương pháp so khớp dựa trên sự tương xứng hình dạng cổ điển là phương pháp *Hausdorff distance*[3]. Nó thường dùng để xác định vị trí của đối tượng trên hình ảnh và đo độ tương đồng giữa các hình.

Ngoài những phương pháp trên, còn có các phương pháp khác như *shape signature*[4, 5, 6], *boundary moments*[7, 8], *elastic matching*[9], *scale space method*[10],...

- Tiếp cận rời rạc: là chia nhỏ hình dạng thành nhiều đoạn dựa trên một tiêu chuẩn đặc biệt, các đoạn được gọi là các hình cơ bản (primitives). Biểu diễn cuối cùng sẽ là một chuỗi ký tự, một biểu đồ (hoặc cây). Độ tương đồng giữa các hình dạng sẽ được đo bằng cách so khớp các chuỗi hoặc biểu đồ đó. Các phương pháp phổ biến dựa trên tiếp cận rời rạc phổ biến gồm *polygonal approximation*, *curvature decomposition* và *curve fitting*[11].

**Biểu diễn hình dạng dựa trên vùng.** Trong kỹ thuật này, tất cả các pixel nằm trên vùng hình dạng của đối tượng được xử lý để thu được một dạng biểu diễn

hình dạng thay vì chỉ dựa trên thông tin đường biên như các phương pháp dựa trên đường viền đã trình bày ở mục trước. Các phương pháp dựa trên vùng sử dụng các bộ mô tả mômen để mô tả hình dạng như *geometric moment invariants*[12], *algebraic moment invariants*[13, 14], *orthogonal moments*[15]. Ngoài ra còn một số phương pháp mô tả dựa trên vùng khác như grid method, shape matrix, convex hull và media axis. Tương tự như các phương pháp dựa trên đường viền, các hình dạng dựa trên vùng có thể được tạo thành theo thứ tự bất kỳ và bất biến trước các biến đổi afin. Tuy nhiên, trong công trình [16], tác giả Meier đã chỉ ra rằng những mômen đại số bất biến cho kết quả tất tốt hoặc rất xấu trên từng đối tượng truy vấn. Chúng cho kết quả tốt trên các đối tượng mà các pixel phân tán chứ không phải là viền của đối tượng.

### 2.2.1.2 Đặc trưng texture

Tương tự như hình dạng, texture là một trong những đặc trưng quan trọng để mô tả và nhận dạng hình ảnh. Điều này đã được chứng minh qua rất nhiều công trình nghiên cứu về phân tích texture của hình ảnh[17, 18, 19, 20].

Texture có thể dễ dàng nhận biết trên hình ảnh. Ví dụ các texture thường thấy như cát, kim loại, gỗ, v.v... Tuy nhiên để một định nghĩa chuẩn cho nó thì không hề dễ dàng. Không giống như màu sắc, texture rất khó để phân tích nếu chỉ dựa trên giá trị của các pixel độc lập bởi mà phải đặt trong mối quan hệ với các pixel xung quanh. Do đó ta có thể nêu ra được các thuộc tính của texture. Theo công trình [21], texture có các thuộc tính như độ thô ráp, độ tương phản, hướng, tính đều đặn, tính thô,...

Việc rút trích đặc trưng về texture có thể phân thành các hướng tiếp cận sau:  
**Phương pháp thống kê.** Đây là một trong những cách truyền thống để phân tích được sự phân bố của độ xám trong không gian, chẳng hạn như tính toán xác suất xuất hiện của các giá trị độ xám ở các khoảng cách và hướng khác nhau. Việc thống kê có thể được thực hiện trên những giá trị của các pixel độc lập hoặc trên giá trị các cặp pixel[17]. Các phương pháp biểu diễn texture bằng histogram cũng sử dụng những thông tin thống kê về texture. Một trong những phương pháp thống kê phổ biến nhất hiện nay là *co-occurrence matrix*[22].

**Phương pháp hình học.** Phương pháp này phân tích texture bằng những thành

phần gốc tạo nên texture. Sự phân tích đó dựa trên các thuộc tính hình học như kích cỡ, hình dáng, diện tích và độ dài. Sau khi xác định được những thành phần gốc đó trên hình ảnh, các quy luật sắp đặt sẽ được rút trích từ đó[23]. Tuy nhiên, các này không thể áp dụng cho những texture từ tự nhiên bởi vì những thành phần gốc và các luật sắp đặt có thể không phổ biến.

**Phương pháp dựa trên mô hình.** Phương pháp này ứng dụng kỹ thuật xây dựng mô hình cho hình ảnh để mô tả và tổng hợp texture. Các thông số của mô hình sẽ lưu lại các thành phần trực giác của texture[17]. Ví dụ, các thành phần của texture có thể được mô hình như sau:các chấm đen và chấm sáng, sự dịch chuyển ngang hoặc dọc, các góc, các đường, v.v... Các bộ mô tả thuộc phương pháp này làm việc tốt với những texture phổ biến. Bộ mô tả *local binary pattern* là một ví dụ cho bộ mô tả dựa trên mô hình.

**Phương pháp xử lý tín hiệu.** Các phương pháp xử lý tín hiệu mô tả texture bằng cách sử dụng các bộ lọc trên toàn hình ảnh. Cả bộ lọc theo không gian và tần suất cũng có thể được sử dụng. Các bộ mô tả dựa trên wavelet và Gabor đều thuộc phương pháp này. Ví dụ, *homogeneous texture descriptor*[24, 18] là một bộ mô tả như vậy.

### 2.2.1.3 Đặc trưng màu sắc

Một trong những thuộc tính quan trọng nhất của đối tượng có thể nhìn thấy bằng mắt thường là màu sắc bởi vậy nó là một trong những thuộc tính quan trọng được sử dụng trong các hệ thống truy vấn dựa trên nội dung ảnh.

Các hướng tiếp cận về đặc trưng màu sắc có thể chia làm 3 dạng:

**Hướng tiếp cận tổng thể.** Hướng tiếp cận này xem xét các đặc trưng dưới góc độ tổng thể, do đó trong quá trình rút trích không hề có sự chia nhỏ hay tiền xử lý. Các bộ mô tả theo hướng này thường sử dụng những thuật toán đơn giản và nhanh gọn để rút trích các vector đặc trưng. Tuy nhiên, các thông tin về sự phân bố không gian của các màu sắc bị bỏ qua nên làm cho các bộ mô tả đó có ít giá trị trong việc phân tách hình ảnh. Có rất nhiều hướng tiếp cận tổng quan sinh ra các histogram hay các vector đặc trưng như *global color histogram*[25] và *cumulative global color histogram*[26].

**Hướng tiếp cận dựa trên các vùng có kích thước cố định.** Hướng tiếp cận

này chia hình ảnh thành các ô với kích thước cố định và rút trích thông tin màu sắc từ mỗi ô một cách độc lập. Các bộ mô tả dựa trên phương pháp này càng mã hóa nhiều thông tin ảnh thì chi phí sẽ càng tăng. Một vài bộ mô tả dựa trên hướng tiếp cận này như: *local color histogram*[25] và *cell/color histogram*[27].

**Hướng tiếp cận dựa trên sự phân đoạn.** Hướng tiếp cận này chia hình ảnh thành nhiều vùng, các vùng có thể có kích cỡ và số lượng khác nhau với các hình khác nhau. Quá trình phân chia này được thực hiện bởi một thuật toán chia đoạn hoặc gom cụm, điều này làm gia tăng chi phí tính toán cho quá trình rút trích đặc trưng. Một loại khác của sự phân đoạn là phân lớp các pixel trước khi rút trích đặc trưng. Các bộ mô tả dựa trên hướng tiếp cận này thường cho kết quả tốt hơn mặc dù chi phí tính toán cao hơn. Một số ví dụ về các bộ descriptor dựa trên phương pháp chia đoạn như *color-based clustering*[27] và *dominant-color*[28, 29].

Mặc dù các đặc trưng toàn cục kể trên đều là những đặc trưng quan trọng nhất để nhận diện hình ảnh nhưng các hướng tiếp cận dựa trên các đặc trưng toàn cục vẫn chưa đạt được kết quả cao trong phân lớp và truy vấn ảnh. Một trong những hướng tiếp cận thu hút được nhiều sự chú ý và đạt được nhiều kết quả triển vọng trong những năm gần đây đó là tiếp cận dựa trên đặc trưng cục bộ. Chúng tôi sẽ trình bày chi tiết về hướng tiếp cận này ở phần kế tiếp.

### 2.2.2 Đặc trưng cục bộ

Trong những năm gần đây, rất nhiều công trình nghiên cứu đã khai thác các đặc trưng cục bộ để biểu diễn hình ảnh phục vụ cho các bài toán về truy vấn và phân lớp ảnh và đã đạt được những thành quả đáng khích lệ.

Bản chất của hướng tiếp cận này 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 bất biến 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

## Chương 2. Các công trình liên quan

---

tâm và đồng thời loại bỏ những chi tiết không đảm bảo sự bất 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.

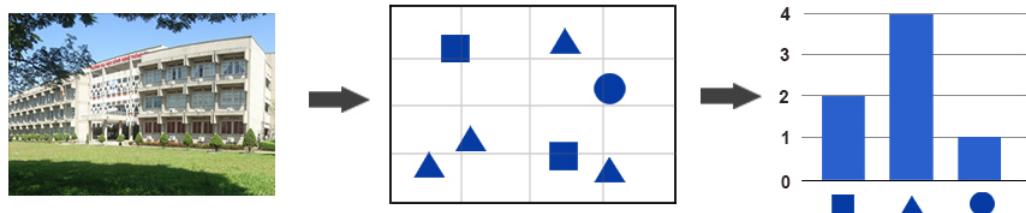
Có rất nhiều phương pháp dò tìm các điểm đặc trưng (feature detector) được đưa ra, trong đó phải kể tới các phương pháp được dùng phổ biến như Difference of Gaussians, DoG [30], Maximally Stable Extremal Regions, MSER [31] và affine invariant detector [32]. Ngoài ra còn có các phương pháp dò tìm được xây dựng để tìm kiếm trong thời gian thực như SURF [33], FAST [34] và BRISK [35].

Sau khi rút trích được các điểm 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 [30], 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 [36], SURF [33], DAISY [37], CONGAS [38], BRIEF [39]. Đặc biệt, bằng việc đề xuất thuật toán RootSIFT được cải tiến từ SIFT, Arandjelovic và Zisserman [40] 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.

## 2.3 Biểu diễn hình ảnh bằng mô hình Bag-of-words

Trong truy vấn hình ảnh, Bag-of-words (BoW) là mô hình biểu diễn hình ảnh được sử dụng phổ biến nhất hiện nay và đạt được những kết quả rất tốt, chẳng hạn như công trình cho kết quả tốt nhất hiện nay của Arandjelovic và Zisserman[40] cũng sử dụng mô hình này. Hình 2.2 minh họa ý tưởng biểu diễn hình ảnh với mô hình BoW.



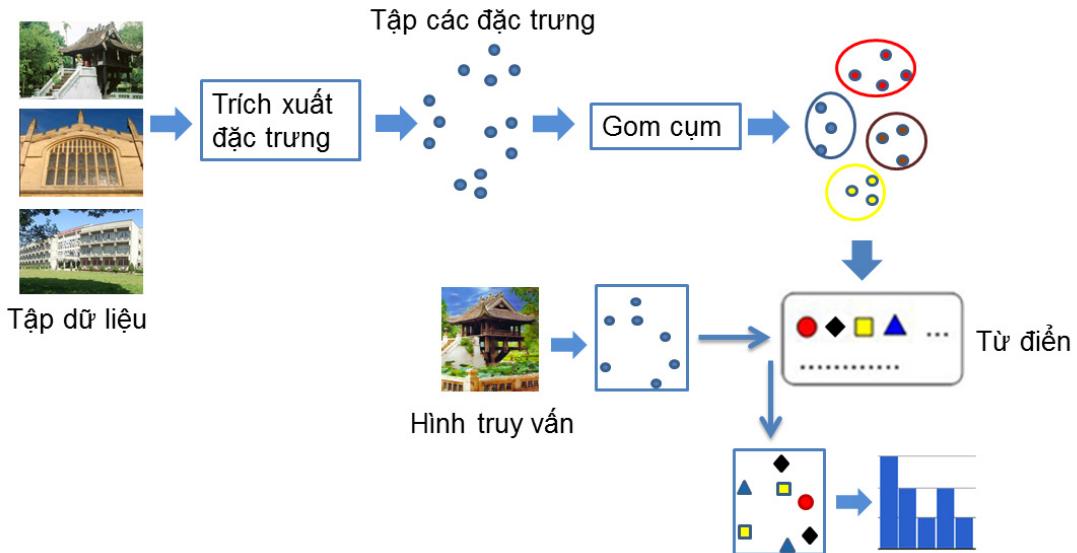
Hình 2.2: Biểu diễn hình ảnh bằng mô hình Bag-of-words.

Mô hình BoW được ứng dụng từ lĩnh vực xử lý văn bản, với ý tưởng biểu diễn một hình ảnh như một văn bản. Cụ thể hơn, BoW biểu diễn các đặc trưng cục bộ bằng các visual word, và từ những visual word sẽ xây dựng được một histogram biểu diễn hình ảnh đó. Như đã giới thiệu trong Mục 2.2.2, một hình ảnh có thể rút trích được 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 visual word từ các đặc trưng này? Hình 2.3 cho thấy các bước xử lý trong mô hình BoW.

Nghiên cứu của Sivic và Zisserman [41] là công trình đầu tiên ứng dụng hướng tiếp cận trong xử lý văn bản vào truy vấn ảnh<sup>1</sup>. Trong công trình này tác giả đã giới thiệu khái niệm visual word, được tạo ra bằng cách sử dụng thuật toán gom cụm K-Means để gom cụm các đặc trưng cục bộ. Hình 2.4 cho thấy một ví

<sup>1</sup>Mục 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

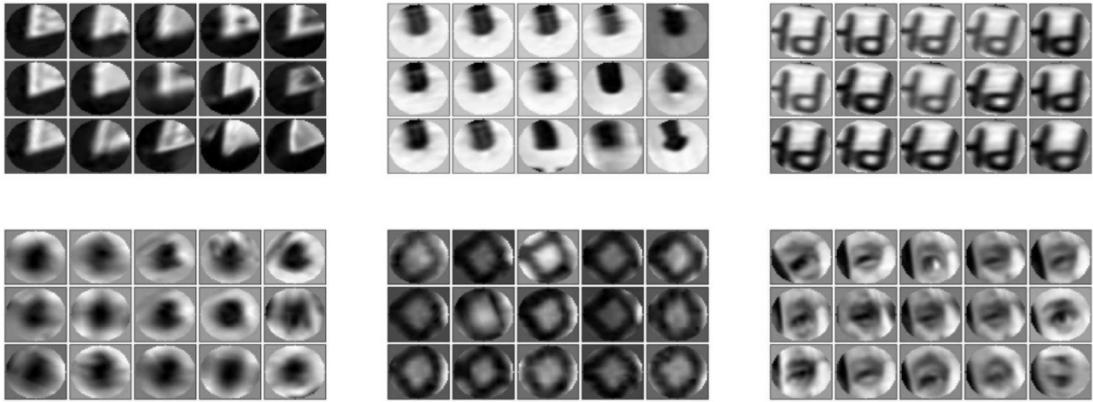
## Chương 2. Các công trình liên quan



Hình 2.3: Mô hình Bag-of-words.

dụ về các visual word. Tương tự như trong truy vấn văn bản, hình ảnh sẽ được rút trích các đặc trưng cục bộ rồi tiệm hành gom cụm các đặc trưng này sẽ thu được các visual word. Tập các visual word này được gọi là một bộ từ điển. Khi một hình ảnh được đưa vào từ điển, mỗi đặc trưng của nó sẽ được đánh chỉ số theo visual word có khoảng cách gần nhất với đặc trưng đó.

Thí nghiệm trong công trình của Sivic và Zisserman [41] được tiến hành trên 4000 ảnh (frame) được lấy từ video, sử dụng 10,000 visual word được gom cụm từ các đặc trưng cục bộ của các hình ảnh đó. Trong thực tế, để truy vấn ảnh trên những tập dữ liệu lớn, để cho kết quả tốt thì số lượng visual word không thể vào khoảng 10,000 từ mà phải lên tới hàng triệu từ [43]. 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 kích cỡ của visual word và số tập của bộ mô tả huấn luyện (training descriptor set). 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 nếu dùng K-Means cho bài toán này chi phí tính toán sẽ vô cùng lớn. Nister và Stewenius [44] 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 Hierarchical K-Means (HKM). Để minh họa cho thuật toán này, tác giả đã cho thử nghiệm



Hình 2.4: **Các visual word.** 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 visual word. Hình ảnh được lấy từ bài báo [42].

trên bộ ảnh gồm 1 triệu hình ảnh. Không lâu sau đó, Philbin và các đồng nghiệp [43] đã đề xuất một hướng tiếp cận khác dựa trên thuật toán *xấp xỉ K-Means*, Approximate K-Means (AKM). 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 những 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  $O(N_d \log(N_w))$ .

## 2.4 So khớp hình ảnh

Quá trình so khớp hình ảnh phụ thuộc phần lớn vào việc hình ảnh đã được biểu diễn như thế nào. Với mô hình BoW như đã trình bày tại Mục 2.3, mỗi hình ảnh sẽ trở thành một histogram, khoảng cách giữa các histogram này thể hiện độ tương tự giữa các hình ảnh.Thêm vào đó, các histogram này cần được đánh trọng số thích hợp để tăng độ chính xác truy vấn. Cũng giống như mô hình BoW trong xử lý văn bản, *tf-idf*[45] là phương pháp tiêu chuẩn để đánh trọng số. *Tf-idf* trong xử lý văn bản được định nghĩa như sau:

Giả sử với bộ từ điển gồm  $k$  từ, mỗi từ sẽ được biểu diễn bằng một vector k chiều  $V_i = (t_1, t_2, \dots, t_i, \dots, t_k)^T$ , với:

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i} \quad (2.1)$$

trong đó,  $n_{id}$  là số lần xuất hiện của từ  $i$  trong văn bản  $d$ ,  $n_d$  là tổng số từ trong văn bản  $d$ ,  $n_i$  là số lần xuất hiện của từ  $i$  trong toàn bộ từ điển và  $N$  là tổng số văn bản trong từ điển. Phương pháp đánh trọng số này gồm hai thành phần chính là *tf* (*term frequency*) là tần số xuất hiện của từ trong văn bản ( $\frac{n_{id}}{n_d}$ ) thể hiện mức độ quan trọng của từ đó trong văn bản và *idf* (*inverse document frequency*) thể hiện mức độ ảnh hưởng của từ đó tới toàn bộ dữ liệu. Một từ có thể có mức độ quan trọng cao nếu nó xuất hiện nhiều lần trong một văn bản, nhưng nếu nó xuất hiện trong hầu hết các văn bản thì nó lại không có ý nghĩa nhiều để phân loại hoặc so sánh các văn bản với nhau.

Trong truy vấn và phân loại hình ảnh sử dụng mô hình BoW, tf-idf được sử dụng tương tự như trên để đánh trọng số cho các histogram biểu diễn hình ảnh, với các từ là các visual word và các văn bản là các hình ảnh. Sau đó, khoảng cách giữa các histogram được tính bằng một trong các độ đo phổ biến:  $L_1$  distance[46], Euclidean distance ( $L_2$ )[47] và histogram intersection (HI)[48].

### L1 distance

$L_1$  distance là độ đo khoảng cách cơ bản giữa hai véc tơ. Khoảng cách  $d$  giữa hai véc tơ  $p, q$  được tính như sau:

$$d(p, q) = d(q, p) = \|p - q\| = \sum_{i=1}^n |p_i - q_i| \quad (2.2)$$

Trong đó  $p = (p_1, p_2, p_3, \dots, p_n)$  và  $q = (q_1, q_2, q_3, \dots, q_n)$

### L2 distance

$L_2$  distance là độ đo khoảng cách được sử dụng khá phổ biến và cho kết quả tốt trong nhiều trường hợp, được tính như sau:

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + \dots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (2.3)$$

trong đó  $p, q$  là các véc  $n$  chiều,  $p = (p_1, p_2, \dots, p_n)$  và  $q = (q_1, q_2, \dots, q_n)$ ,  $d(p, q)$  là khoảng cách giữa  $p$  và  $q$ .

### Histogram Intersection

Được đưa ra năm 2010 bởi Erkang Cheng và các đồng nghiệp[48], histogram intersection (HI) cũng thể hiện được hiệu quả của nó trong những trường hợp nhất định. Giá trị  $S_{HI}$  thể hiện mức độ tương đồng giữa hai histogram  $h_1$  và  $h_2$  được định nghĩa như sau:

$$S_{HI}(h_1, h_2) = \sum_{i=1}^n \min(h_1(i), h_2(i)) \quad (2.4)$$

Tác giả cũng đưa ra phương pháp Normalized HI (NHI):

$$S_{NHI}(h_1, h_2) = \sum_{i=1}^n \frac{\min(h_1(i), h_2(i))}{h_1(i) + h_2(i)} \quad (2.5)$$

cũng cho kết quả khả quan trong một số trường hợp khác.

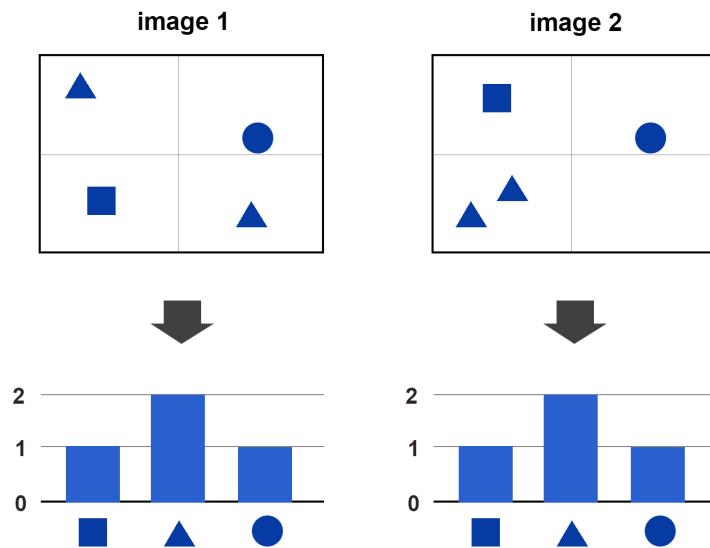
Với ba độ đo trên ta có thể thấy,  $L1$  và  $L2$  thể hiện cả mức độ giống nhau và khác nhau giữa hai histogram, trong khi đó histogram intersection quan tâm chủ yếu tới mức độ giống nhau giữa hai histogram. Việc sử dụng độ đo nào cho thích hợp tùy thuộc vào các trường hợp cụ thể, và nó còn liên quan tới nhiều vấn đề khác như kích thước từ điển ảnh hưởng tới kích thước mỗi histogram, mức độ thừa hay nhiễu của các histogram...

Trong truy vấn hình ảnh, quá trình so khớp được sử dụng khi có một hình truy vấn được đưa vào, nó sẽ được so khớp với các hình trong bộ dữ liệu. Để tăng tốc quá trình truy vấn, phương pháp chỉ mục ngược (inverted index) được

sử dụng phô biến và cho thấy hiệu quả cao giúp cải thiện hiệu suất truy vấn.

## 2.5 Sử dụng thông tin về sự phân bố trong không gian ảnh của các đặc trưng

Mặc dù đạt được những kết quả rất đáng chú ý nhưng mô hình BoW cơ bản vẫn bị giới hạn về độ chính xác do bỏ qua một thông tin quan trọng, đó là sự phân bố về không gian của các visual word. Do đó các đặc trưng cục bộ được xử lý một cách rời rạc, không liên quan tới nhau. Hình 2.5 minh họa cho trường hợp các ảnh khác nhau được biểu diễn như nhau nếu không xem xét sự phân bố về không gian của các visual word.



Hình 2.5: **Bỏ qua thông tin về sự phân bố trong không gian của các visual word trong mô hình BoW.** Sau khi được biểu diễn bằng mô hình BoW, hai hình ảnh trên được coi như giống nhau hoàn toàn trong khi chúng khác nhau do các visual word nằm ở các vị trí khác nhau.

Để giải quyết vấn đề trên, rất nhiều công trình nghiên cứu đã được đưa ra. Phần lớn các hướng tiếp cận 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 điểm đặc trưng

cục bộ. Trong mục 2.5.1 chúng tôi sẽ trình bày về các phương pháp dựa trên đặc trưng hình học. Hướng tiếp cận còn lại sẽ được trình bày chi tiết trong mục 2.5.2.

### 2.5.1 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 và Zisserman [41] đã đo đạc 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 đạc sự nhất quán không gian cục bộ trong so khớp hình ảnh cũng được đề cập tới trước đó trong các công trình như [49] và [50].

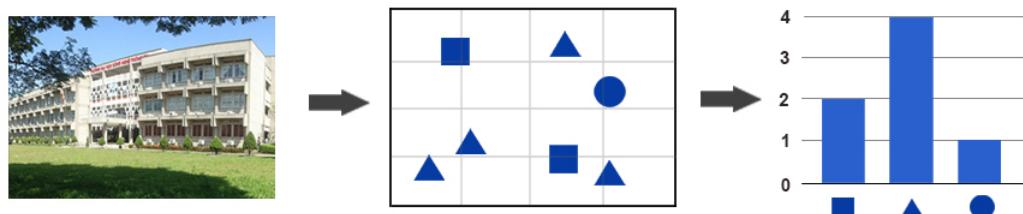
Trong một công trình nghiên cứu [43], tác giả sử dụng thuật toán RANSAC [51] để kiểm tra sự nhất quán hình học giữa các đặc trưng cục bộ trùng khớp. RANSAC là một trong những phương pháp phổ biến nhất cho hậu xử lý toàn cục trên hình ảnh. Đặc biệt, trong một công trình khác, Zhang và các đồng nghiệp [52] 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, công trình [53] và [54] 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 mô hình BoW 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.

### 2.5.2 Các hướng tiếp cận dựa trên thông tin không gian của các điểm 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 [55] đã giới thiệu một phương pháp nền tảng, được bắt nguồn từ ý tưởng *so khớp phân cấp* (pyramid matching) của Grauman và Darrell [56], 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 BoW 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.



Hình 2.6: Phương pháp Spatial Pyramid Matching.

## **2.6 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 BoW 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 BoW 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 truy vấn nhưng lại cần chi phí tính toán cao, tốc độ phản hồi chậm.

Vì vậy, với những hạn chế trên, chúng tôi đề xuất một phương pháp tập trung vào việc cân bằng độ chính xác truy vấn và tốc độ phản hồi.

# Chương 3

## Phương pháp đề xuất

Rất nhiều công trình được đưa ra để giải quyết bài toán truy vấn ảnh như đã trình bày trong mục 2.2 và 2.3. Tuy nhiên các phương pháp trên còn tồn tại những hạn chế mà một trong số đó là việc bỏ qua thông tin không gian ảnh. Trong khi đó cũng có nhiều phương pháp được đưa ra trong những năm gần đây để giải quyết vấn đề này được giới thiệu trong mục 2.5, nhưng cũng chưa thực sự hiệu quả. Chẳng hạn như phương pháp spatial pyramid matching[55] sử dụng thông tin không gian ảnh để cải thiện độ chính xác nhưng chi phí tính toán lại rất lớn, thời gian phản hồi chậm nên khó có thể đáp ứng những bộ dữ liệu ngày càng lớn như hiện nay.

Trong chương này, chúng tôi sẽ trình bày phương pháp đề xuất tích hợp thông tin không gian ảnh vào chỉ mục ngược, với mục tiêu cân bằng độ chính xác truy vấn và thời gian phản hồi. Trước tiên chúng tôi nhắc lại những công trình khởi nguồn ý tưởng cho phương pháp đề xuất, sau đó trình bày chi tiết về ý tưởng cải tiến và phương pháp đề xuất.

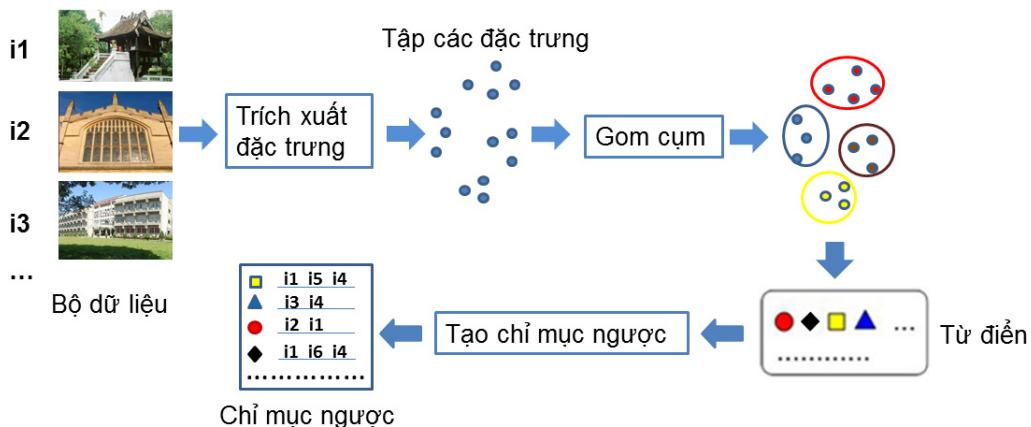
### 3.1 Chỉ mục ngược trong truy vấn hình ảnh

Như đã được giới thiệu sơ lược trong mục 2.4, chỉ mục ngược (inverted index) là phương pháp phổ dùng để tăng 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 khác, chỉ mục ngược là một cấu trúc dữ liệu chủ yếu bao gồm 2 trường là khóa và giá trị.

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

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 trên từng hình ảnh trong bộ dữ liệu, 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 visual word, mỗi visual word 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 visual word tương ứng. Quá trình tạo tập chỉ mục ngược (inverted file) được minh họa trong Hình 3.1.



Hình 3.1: Quá trình tạo chỉ mục ngược.

Với cấu trúc đơn giản nhưng chỉ mục người lại cho thấy hiệu quả to lớn của nó. Chẳng hạn khi có một hình ảnh truy vấn được đưa vào, có thể dễ dàng tìm tất cả các hình ảnh liên quan tới nó mà không cần quan tâm tới những hình ảnh không liên quan khác, hoặc trong quá trình tính trọng số tf-idf cho một visual word được trình bày trong mục 2.4 cũng có thể nhanh chóng tìm tất cả những

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

hình ảnh chứa visual words đó mà không phải truy xuất vào dữ liệu của từng hình ảnh.

Với việc sử dụng chỉ mục ngược, tốc độ truy vấn được tăng lên đáng kể. Cũng chính vì vậy, chúng tôi đề xuất ý tưởng tích hợp thông tin không gian ảnh vào chỉ mục ngược nhằm tận dụng tính đơn giản nhưng hiệu quả của nó.

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

Phương pháp chúng tôi đề xuất bắt nguồn từ ý tưởng tận dụng tính hiệu quả của chỉ mục ngược, kết hợp với việc sử dụng thông tin không gian ảnh trong công trình nghiên cứu của Lazebnik và các đồng nghiệp[55]. Để tận dụng thông tin về sự phân bố trong không gian của các visual word, chúng tôi chia các hình ảnh thành nhiều ô tại nhiều cấp độ khác nhau. Với mỗi ô tại mỗi cấp độ, chúng tôi lấy được danh sách các visual word nằm trong ô đó của tất cả các hình ảnh, từ đó xây dựng được một chỉ mục ngược. Mỗi chỉ mục ngược tương ứng với một ô trong không gian ảnh, lưu trữ tất cả các visual word nằm trong ô đó, đồng thời lưu trữ chỉ số của các hình ảnh có chứa các visual word này trong ô đó. Chẳng hạn tại cấp độ thứ hai, hình ảnh được chia làm bốn phần, mỗi ô là một góc phần tư. Với mỗi góc phần tư trong mỗi hình ảnh sẽ tìm được các visual word nằm trong góc này, và khi duyệt qua toàn bộ bộ dữ liệu ta sẽ xây dựng được một chỉ mục ngược cho góc phần tư này. Tại quá trình truy vấn, khi một hình ảnh được đưa vào, nó sẽ được chia ra như trong quá trình tạo chỉ mục ngược đã thực hiện trước đó. Với mỗi ô tại mỗi cấp độ, các visual word sẽ được lọc ra, từ đó đưa vào chỉ mục ngược tương ứng để tìm ra tất cả những hình ảnh cũng chứa các visual word này tại ô này.

Cụ thể hơn, theo nghiên cứu của Lazebnik, hình ảnh được chia thành 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  $1 \times 1$ ; cấp 1 là  $2 \times 2$ . 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

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

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 nhau rồi rồi được ghép nối với nhau để tạo thành một véc tơ đặc trưng biểu diễn cho hình ảnh, độ dài của véc tơ này sẽ tăng  $\frac{1}{3} \times (4^{L+1} - 1)$  lần. Bằng cách biểu diễn như vậy, các hình ảnh có sự phân bố các visual word tương tự nhau sẽ được biểu diễn bằng những biểu đồ ghép nối gần giống nhau.

Trong phương pháp đề xuất này, chúng tôi chỉ áp dụng cách chia hình thành không gian phân cấp theo nghiên cứu trên, sau đó kết quả của quá trình chia hình được sử dụng một cách hoàn toàn khác. Phương pháp của Lazebnik cho thấy độ phức tạp lớn trong quá trình so khớp, vì vậy nó thường sử dụng để hậu xử lý kết quả. Còn trong phương pháp của chúng tôi, thông tin không gian ảnh được tích hợp vào chỉ mục ngược, được sử dụng ngay trong bước tiền xử lý, không làm tăng thêm nhiều chi phí tính toán.

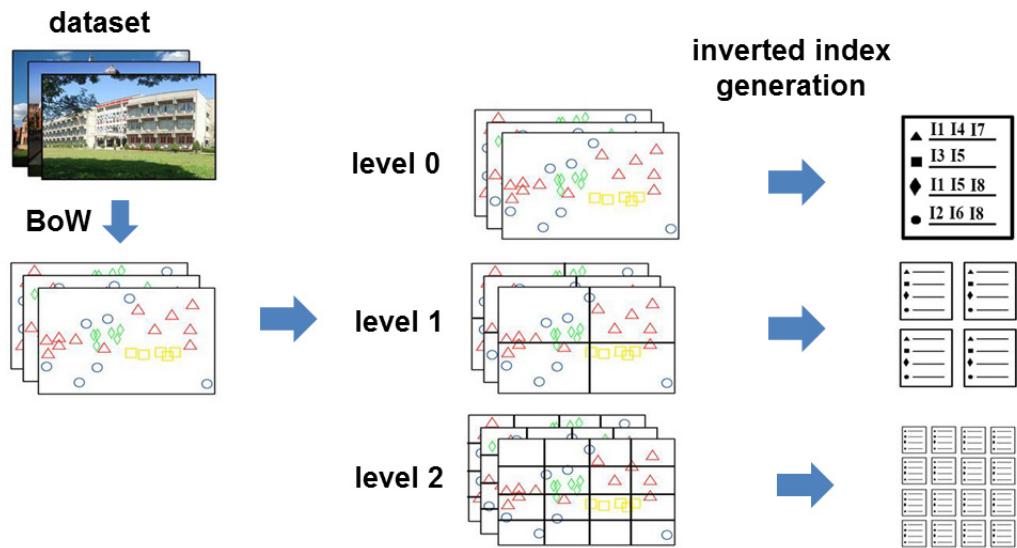
Để sử dụng thông tin không gian ảnh đã được tích hợp vào chỉ mục ngược, chúng tôi đề xuất sử dụng một kỹ thuật xếp hạng được gọi là *bầu chọn* (voting). Kỹ thuật được định nghĩa như sau: trong quá trình truy vấn, các đặc trưng sẽ được rút trích từ hình ảnh truy vấn, sau đó từ các đặc trưng ta sẽ lấy được các visual word bằng cách sử dụng từ điển sau đó tra cứu trong tập chỉ mục ngược để lấy được các hình ảnh ứng viên. Những hình ảnh nào có số lượng visual word trùng với các visual word trong hình ảnh truy vấn càng nhiều thì sẽ càng được xếp hạng cao hơn trong danh sách kết quả truy vấn trả về.

Với kỹ thuật bầu chọn trong phương pháp đề xuất, chúng tôi 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 visual word giống nhau trong cùng một ô sẽ nhận được nhiều lượt bầu chọn hơn so với hai hình ảnh có các visual word 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 để thể hiện mức độ quan trọng của các cấp độ chia hình ảnh khác nhau. Chúng tôi đánh trọng số giống phương pháp của Lazebnik[55]. Trọng số tại cấp thứ nhất ( $l = 0$ ) sẽ là  $\frac{1}{2^L}$ , tại các cấp  $l$  tiếp theo sẽ là  $\frac{1}{2^{L-l+1}}$ . Đây là phương pháp đánh trọng số được đề xuất bởi K. Grauman and T. Darrell[56].

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 chỉ mục ngược khác nhau nhưng các tập vẫn giữ được cấu trúc căn bản của chỉ mục ngược. Mỗi tập sẽ dùng để

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

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 chỉ mục ngược sẽ là  $\frac{1}{3}(4^{L+1} - 1)$  và mỗi cấp độ sẽ có  $2^l \times 2^l$  tập chỉ mục ngược với  $0 \leq l \leq L$ . Hình 3.2 mô tả khái quát cho phương pháp được đề xuất.

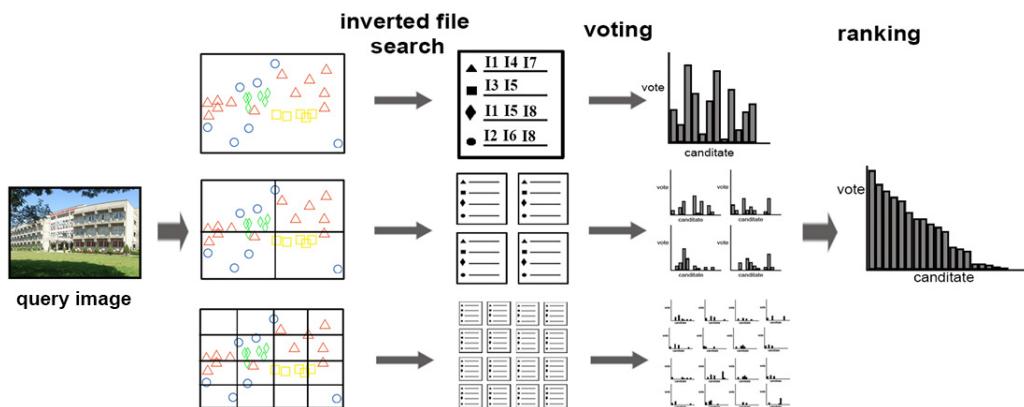


Hình 3.2: Khái quát về phương pháp đề xuất.

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 visual word (từ điển). Mỗi hình ảnh sẽ chứa một tập các visual word. 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 visual word đó có thuộc ô đang xét hay không. Tất cả các visual word được tìm thấy trong mỗi ô sẽ được thu thập lại. Tiếp theo, tập hợp của các visual word được tìm thấy trong mỗi ô của các hình ảnh sẽ được dùng để sinh ra một tập chỉ mục ngược tương ứng với ô đó. Số lượng tập chỉ mục ngược được sinh ra bằng với tổng số ô của không gian phân cấp.

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

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 visual word tương ứng. Dựa vào vị trí của các visual word 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 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 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 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.



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

Hiệu quả của phương pháp đề xuất sẽ được chứng minh bằng những thí nghiệm được trình bày trong phần tiếp theo.

# Chương 4

## Thực nghiệm và đánh giá kết quả

Để đánh giá hiệu suất của một hệ thống truy vấn ảnh, ta cần cài đặt và thử nghiệm với những quy trình đánh giá chuẩn. Đồng thời so sánh nó với các hệ thống khác trong cùng một điều kiện thí nghiệm.

Với ý tưởng như đã trình bày trong chương trước, trong Chương này chúng tôi sẽ mô tả chi tiết cách cài đặt thí nghiệm cũng như quy trình đánh giá hiệu suất của hệ thống để xuất đồng thời so sánh kết quả với các phương pháp khác. Trước tiên, chi tiết về các bộ dữ liệu và phương pháp dùng để đánh giá sẽ được trình bày một cách chi tiết trong mục 4.1. Cách cài đặt các phương pháp cơ sở cũng như phương pháp đề xuất sẽ được mô tả trong mục 4.2. Và cuối cùng trong mục 4.3, chúng tôi sẽ đánh giá phương pháp đề xuất và so sánh với các phương pháp khác dựa trên kết quả thí nghiệm thu được để đưa ra kết luận.

### 4.1 Các bộ dữ liệu và phương thức đánh giá

Mục này 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. Thực nghiệm và đánh giá kết quả

### 4.1.1 Các bộ dữ liệu

#### 4.1.1.1 Oxford 5K

Bộ dữ liệu Oxford 5K được xây dựng bởi Philbin và các đồng nghiệp [43], bao gồm 11 Oxford "landmark"<sup>1</sup> 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 5,063 hình ảnh chất lượng cao ( $1366 \times 768$ ).

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 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 ( $1366 \times 768$ ) 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) [57].

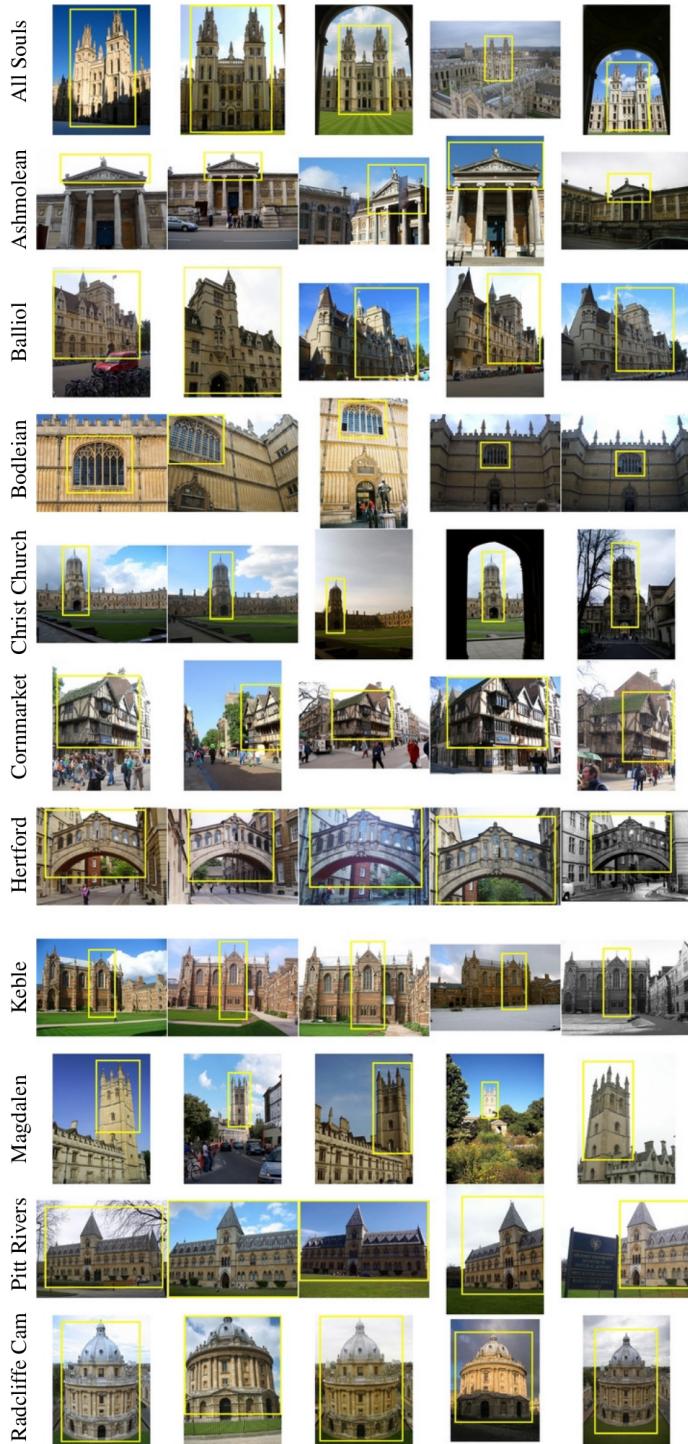
Paris 6K được đánh giá là một bộ dữ liệu hoàn toàn độc lập so với Oxford 5K và thường được dùng để kiểm tra các tác động của việc tính toán visual word trong khi Oxford 5K thường được dùng để kiểm tra hiệu suất.

#### 4.1.1.3 Oxford 5K+100K

Bộ dữ liệu Oxford 5K+100K là bộ dữ liệu được tổng hợp từ hai bộ dữ liệu là Oxford Building 5K và Oxford 100K. Bộ dữ liệu này gồm 105,134 hình ảnh chất lượng cao (5,063 hình từ bộ Oxford Building 5K và 100,071 hình ảnh từ bộ Oxford

<sup>1</sup>landmark ở đây có nghĩa là một góc nhìn/góc chụp đặc biệt của một tòa nhà

#### 4. Thực nghiệm và đánh giá kết quả



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 [43].

## 4. Thực nghiệm và đánh giá kết quả

Bộ dữ liệu	Số lượng hình ảnh	Số lượng truy vấn
Oxford 5K	5,063	55
Paris 6K	6,412	55
Oxford 5K+100K	100,071	55

Bảng 4.1: Số hình ảnh trong mỗi bộ dữ liệu và số truy vấn trong tập dữ liệu đánh giá chuẩn tương ứng.

100K). Bộ Oxford Building 5K đã trình bày chi tiết ở trên còn bộ Oxford 100K thì cũng được lấy về từ Flickr bằng cách tìm kiếm với 145 từ khóa phổ biến nhất.

Trong bộ dữ liệu này, 100,071 hình ảnh từ bộ Oxford 100K được sử dụng chủ yếu như là các hình gây nhiễu. 55 query được sử dụng như trong bộ Oxford 5K với cùng tập dữ liệu đánh giá chuẩn.

Các bộ dữ liệu trên được tổng hợp trong bảng [4.1](#).

### 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 Goggles<sup>1</sup> 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

<sup>1</sup>Google 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. Thực nghiệm và đánh giá kết quả

(AP) [43], 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. AP được tính cho từng truy vấn một sau đó ta lấy trung bình cộng của chúng, đó chính là mean Average Precision (mAP) - một con số để đánh giá hiệu suất tổng thể của hệ thống.

Để đo hiệu suất về tốc độ truy vấn của các hệ thống, chúng tôi đo thời gian xử lý một truy vấn tính từ thời điểm sau khi rút trích được các đặc trưng tới khi có danh sách xếp hạng các hình ảnh. Chúng tôi không tính các khoảng thời gian khác như trích xuất đặc trưng, tính visual word vì những phần xử lý này nằm ngoài phương pháp đề xuất.

### 4.2 Cài đặt thí nghiệm

#### 4.2.1 Các phương pháp đánh giá cùng thông số cài đặt

Để đánh giá hiệu suất của từng phương pháp, chúng tôi cài đặt các phương pháp cơ sở và phương pháp đề xuất như sau:

- **Phương pháp cơ sở 1:** Sử dụng mô hình BoW + phương pháp chỉ mục ngược cơ bản với xếp hạng dựa trên bầu chọn (voting).
- **Phương pháp cơ sở 2:** Sử dụng mô hình BoW + phương pháp chỉ mục ngược cơ bản với xếp hạng bằng việc tính toán khoảng cách giữa hình ảnh truy vấn với mỗi hình ảnh ứng viên. Các hình ảnh được biểu diễn bằng mô hình SPM [55].
- **Phương pháp đề xuất:** Sử dụng mô hình BoW + phương pháp chỉ mục ngược được tính hợp thông tin không gian ảnh do chúng tôi đề xuất cho việc lập chỉ mục và xếp hạng.

Dưới đây là chi tiết cài đặt thí nghiệm và các thông số cho mô hình BoW của cả ba phương pháp (Hình 4.2) :

**Phát hiện và mô tả các điểm đặc trưng.** Để phát hiện các điểm đặc trưng cho từng hình ảnh, chúng tôi sử dụng bộ phát hiện đặc trưng Hessian-Affine[58]. Đây là một bộ phát hiện bất biến dùng để thu thập các điểm quan tâm trong hình ảnh. Với mỗi điểm đặc trưng, một vector 128 chiều được tạo ra từ bộ mô tả SIFT. Từ vector này, chúng tôi sẽ tính RootSIFT[40] để đạt được hiệu suất tốt

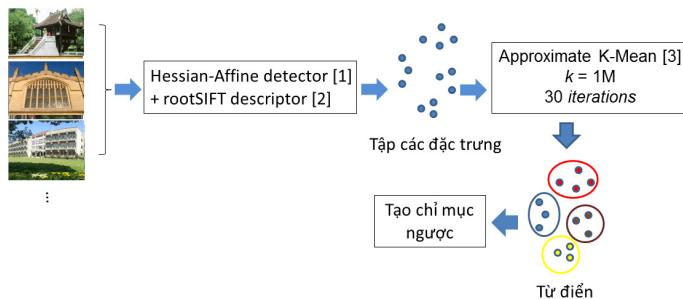
## 4. Thực nghiệm và đánh giá kết quả

hơn.

**Gom cụm các đặc trưng.** Khi gom cụm một tập dữ liệu lớn, ta không thể dùng k-Means bởi vì chi phí tính toán quá lớn. Theo như [43], Approximate K-Means (AKM) có thể được sử dụng để thay thế cho k-Means với một chi phí tính toán chấp nhận được. Ở đây chúng tôi sử dụng AKM để gom cụm thành 1 triệu visual word.

**Tạo chỉ mục ngược.** Như phương pháp đã được trình bày ở trên, số tập chỉ mục ngược sinh ra sẽ bằng với số ô của không gian phân cấp. Mỗi tập chỉ mục sẽ lưu thông tin cho một ô. Tất cả các tập chỉ mục đó sẽ được lưu thành 1 môt tập tin duy nhất.

**Quá trình truy vấn.** Mỗi hình ảnh truy vấn trong tập dữ liệu đánh giá chuẩn sẽ được rút trích các đặc trưng, tính toán ra các visual word từ từ điển rồi sau đó lấy ra danh sách hình ảnh ứng viên từ các tập chỉ mục ngược. Cuối cùng, các hình ảnh ứng viên được xếp hạng bằng phương pháp bầu chọn.



Hình 4.2: Thông số cài đặt thí nghiệm của mô hình BoW

### 4.2.2 Nâng cao hiệu suất của hệ thống

Một hệ thống truy vấn ảnh dựa trên mô hình BoW căn bản phải đối mặt với rất nhiều vấn đề như độ chính xác của quá trình gom cụm các đặc trưng, quá trình so khớp bị "gây nhiễu" bởi các stop word, lựa chọn độ đo khoảng cách phù hợp giữa các histogram,... Để tăng hiệu suất của quá trình truy vấn, dưới đây chúng tôi sẽ đề xuất các kỹ thuật cải tiến và thử nghiệm để so sánh kết quả.

## 4. Thực nghiệm và đánh giá kết quả

Số lần lặp (iterations)	mAP (mean Average Precision)
5	0.6084
10	0.6199
20	0.6249
<b>30</b>	<b>0.6278</b>

Bảng 4.2: So sánh kết quả truy vấn với số lần lặp khác nhau trong thuật toán gom cụm AKM.

### 4.2.2.1 Tăng độ chính xác của quá trình gom cụm

Quá trình gom cụm các đặc trưng tạo thành các visual word càng chính xác thì độ chính xác của quá trình truy vấn càng cao. Như đã trình bày trong mục 2.3, mặc dù đạt được độ chính xác cao nhưng thuật toán k-means đòi hỏi chi phí tính toán vô cùng lớn nên ta không thể áp dụng k-means cho bài toán này. Trong công trình [43], tác giả đã chứng minh lợi thế vượt trội về chi phí tính toán của thuật toán gom cụm AKM so với k-means trong khi độ chính xác gần như nhau. Trong thuật toán k-means, quá trình gom cụm sẽ lặp đi lặp lại cho tới khi đạt được độ chính xác tuyệt đối. Còn AKM sẽ lặp lại quá trình gom cụm với một số lần lặp (iteration) đã được định trước. Độ phức tạp của thuật toán k-means luôn lớn hơn  $O(N_w^2)$  còn của AKM là  $O(N_w \log(N_w))$ . Tuy nhiên, để đạt hiệu suất cao nhất, ta cần phải điều chỉnh số lượng vòng lặp của AKM sao cho phù hợp với lượng tài nguyên giới hạn và đảm bảo kết quả tốt. Trong bảng 4.2, chúng tôi điều chỉnh số lượng vòng lặp (iteration) khác nhau của thuật toán AKM và thử nghiệm trên bộ dữ liệu Oxford 5K để theo dõi sự thay đổi của kết quả trả về.

Bảng 4.2 cho thấy hiệu suất tăng mạnh khi số lần lặp tăng từ 5 lên 10 và giảm một chút khi tăng từ 10 lên 20 mặc dù chi phí tính toán vẫn tăng đáng kể. Khi số lần lặp tăng từ 20 lên 30, chi phí tính toán vẫn tăng nhưng hiệu suất không tăng nhiều. Điều đó cho thấy, càng về sau, khi ta tăng số lần lặp thì chi phí tính toán vẫn tăng nhưng hiệu suất tăng rất ít. Do đó, trong các thí nghiệm của mình, chúng tôi sử dụng thông số *iterations* = 30 để đạt được kết quả tốt nhất và giữ chi phí tính toán ở mức cho phép.

## 4. Thực nghiệm và đánh giá kết quả

Lọc bỏ stop words	mAP (mean Average Precision)
Chưa lọc bỏ	0.6278
<b>Lọc bỏ 5% top</b>	<b>0.6323</b>
Lọc bỏ 10% top	0.6293

Bảng 4.3: Thí nghiệm lọc bỏ các stop words (các visual word có tần số xuất hiện cao nhất trong bộ dữ liệu).

### 4.2.2.2 Lọc bỏ các stop word

Trong truy vấn văn bản, các từ phổ biến và xuất hiện thường xuyên trong văn bản với tần suất cao được gọi là stop word. Các stop word này làm giảm độ chính xác của quá trình truy vấn do không có giá trị nhiều trong việc phân biệt các văn bản. Tương tự, trong mô hình BoW, ta cũng bắt gặp rất nhiều stop word làm giảm độ chính xác của truy vấn, gây tốn không gian lưu trữ và chi phí tính toán. Do đó, chúng tôi tiến hành thêm một bước sàng lọc các visual word bằng cách đếm số lần xuất hiện của các visual word trong các hình ảnh và lọc bỏ một nhóm các visual word có số lần xuất hiện nhiều nhất. Kết quả thí nghiệm được tiến hành trên bộ Oxford 5K và được trình bày trong bảng 4.3.

Kết quả thí nghiệm trong bảng 4.3 cho thấy khi lọc bỏ stop word, độ chính xác của truy vấn tăng lên rõ rệt. Khi lọc bỏ 5% số visual word có tần suất xuất hiện cao nhất, độ chính xác tăng mạnh. Tuy nhiên, nếu ta loại bỏ quá nhiều thì độ chính xác lại giảm xuống do với bộ dữ liệu này, lượng stop word chỉ giới hạn ở mức trong khoảng 5%. Với kết quả trên, chúng tôi sẽ tiến hành tiền xử lý loại bỏ 5% các visual word có tần suất xuất hiện cao nhất ở các thí nghiệm tiếp theo để hệ thống đạt được hiệu suất tốt nhất.

## 4.3 Kết quả thí nghiệm và đánh giá kết quả

Bảng 4.4 thể hiện kết quả chi tiết khi chạy thí nghiệm trên bộ dữ liệu Oxford 5K. Có thể thấy rằng phương pháp cơ sở 1 (sử dụng chỉ mục ngược căn bản và xếp hạng bằng bầu chọn) cho độ chính xác thấp với chỉ số mAP = 0.5678 nhưng tốc độ truy vấn rất nhanh với tổng thời gian truy vấn là 0.0788 giây.

## 4. Thực nghiệm và đánh giá kết quả

Phương pháp (trên Oxford 5K)	mAP	Thời gian truy vấn (55 truy vấn)	Bộ nhớ sử dụng
Phương pháp cơ sở 1	0.5678	0.0788 (s)	68.31MB
Phương pháp cơ sở 2	0.6204	30.1153 (s)	68.31MB
<b>Phương pháp đề xuất (<math>L = 2</math>)</b>	<b>0.5851</b>	<b>0.1651 (s)</b>	<b>481.69MB</b>

Bảng 4.4: Hiệu suất của các phương pháp trên bộ dữ liệu Oxford 5K.

Trong khi đó, với việc tích hợp thông tin không gian ảnh vào trong chỉ mục ngược, phương pháp đề xuất cho độ chính xác mAP = 0.5851. Ở đây chúng tôi sử dụng mô hình không gian phân cấp ở cấp 2 ( $L = 2$ , bao gồm 21 tập chỉ mục ngược). Thời gian truy vấn cho 55 truy vấn từ tập dữ liệu đánh giá chuẩn là 0.1651s (khoảng 3ms cho mỗi truy vấn) cũng không quá chênh lệch so với phương pháp cơ sở 1.

Phương pháp cho độ chính xác cao nhất (mAP = 0.6204) là phương pháp cơ sở 2. Phương pháp này tốn rất nhiều chi phí cho quá trình xếp hạng. Để xếp hạng các ứng viên, phương pháp này tính và so sánh khoảng cách L2 (khoảng cách Euclidean) giữa hình ảnh truy vấn và từng hình ảnh ứng viên, các hình ảnh này được biểu diễn bằng mô hình SPM. Do đó, phương pháp này cho thời gian truy vấn rất lâu, **gấp khoảng 182 lần** so với phương pháp đề xuất.

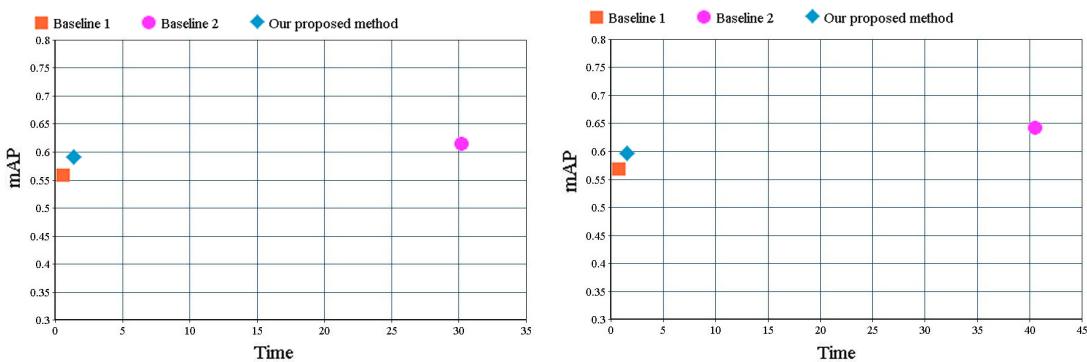
Kết quả trên cho thấy phương pháp của chúng tôi đã cân bằng được độ chính xác và thời gian truy vấn so với các phương pháp khác. Ta có thể thấy kết quả tương tự khi thử nghiệm với bộ Paris 6K. Kết quả được thể hiện trong [Bảng 4.5](#). Biểu đồ trong [Hình 4.3](#) cho thấy sự so sánh tương quan giữa các phương pháp khi thử nghiệm trên bộ dữ liệu Oxford 5K.

Để đánh giá các phương pháp trong điều kiện của các yêu cầu thực tế, các thí nghiệm cần được tiến hành với những bộ dữ liệu có kích thước lớn hơn. Do đó trong thí nghiệm này chúng tôi quyết định sử dụng bộ Oxford 5K+100K để đo đặc hiệu suất của các phương pháp. Tuy nhiên, khi thí nghiệm trên những bộ dữ liệu lớn như vậy, vấn đề lớn nhất phải giải quyết là vấn đề về bộ nhớ. Ví dụ như với bộ Oxford 5K+100K này, chúng tôi rút trích được 294,910,315 vector đặc

#### 4. Thực nghiệm và đánh giá kết quả

Phương pháp (trên Paris 6K)	mAP	Thời gian truy vấn (55 truy vấn)	Bộ nhớ sử dụng
Phương pháp cơ sở 1	0.5762	0.1137 (s)	80.37MB
Phương pháp cơ sở 2	0.6421	40.5526 (s)	80.37MB
<b>Phương pháp đề xuất (<math>L = 2</math>)</b>	<b>0.5967</b>	<b>0.2158 (s)</b>	<b>519.10MB</b>

Bảng 4.5: Hiệu suất của các phương pháp trên bộ dữ liệu Paris 6K.



Hình 4.3: Biểu đồ so sánh độ chính xác và thời gian truy vấn giữa các phương pháp trên bộ Oxford 5K (nên trái) và bộ Paris 6K (bên phải).

#### 4. Thực nghiệm và đánh giá kết quả

---

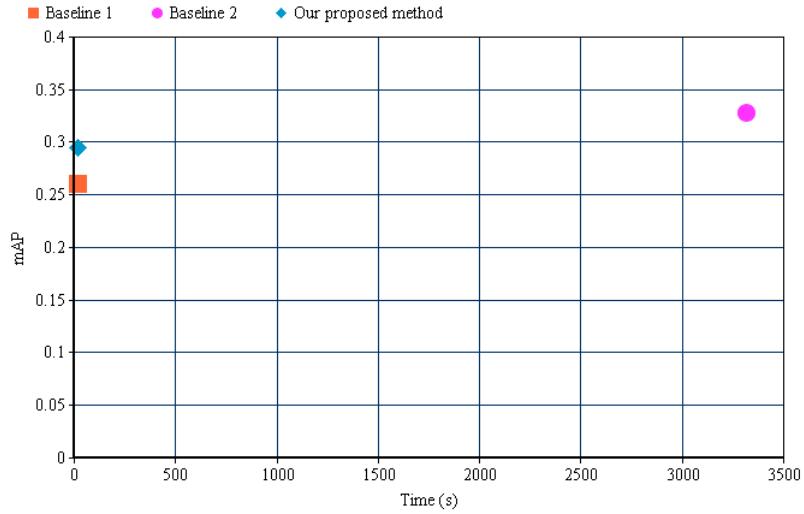
Phương pháp (trên Oxford 5K+100K)	mAP	Thời gian truy vấn (55 truy vấn)	Bộ nhớ sử dụng
Phương pháp cơ sở 1	0.2601	16.1319 (s)	364.14MB
Phương pháp cơ sở 2	0.3279	3315.02 (s)	364.14MB
<b>Phương pháp đề xuất (<math>L = 2</math>)</b>	<b>0.2950</b>	<b>18.4219 (s)</b>	<b>1,369.88MB</b>

Bảng 4.6: Hiệu suất của các phương pháp trên bộ dữ liệu Oxford 5K+100K.

trưng 128 chiều tức là chiếm khoảng 140,6GB bộ nhớ. Con số này vượt xa khả năng về bộ nhớ mà chúng tôi có. Vì thế việc gom cụm những đặc trưng này để lấy được các visual word là một điều không thể. Do đó, chúng tôi chấp nhận hi sinh một phần độ chính xác để có thể tiến hành thí nghiệm trên bộ dữ liệu này. Cụ thể trong thí nghiệm với bộ Oxford 5K+100K, chúng tôi tiến hành thu nhỏ kích cỡ của hình chỉ còn 50% kích cỡ ban đầu trước khi tiến hành thí nghiệm. Đồng thời, sau khi rút trích được các đặc trưng, chúng tôi sẽ lấy ngẫu nhiên  $\frac{1}{3}$  số lượng đặc trưng để sử dụng cho thí nghiệm. Các thông số còn lại đều được sử dụng giống với các thí nghiệm trước. Vì vậy, tuy kết quả có thể bị giảm đi nhưng tương quan giữa các phương pháp không thay đổi, vẫn có thể so sánh các phương pháp với nhau. Kết quả thí nghiệm trên bộ dữ liệu này được thể hiện trong [Bảng 4.6](#). Có thể thấy rằng, phương pháp do nhóm đề xuất vẫn giữ được sự cân bằng giữa độ chính xác và thời gian truy vấn. Biểu đồ trong [Hình 4.4](#) cho thấy sự so sánh hiệu suất giữa ba phương pháp trên bộ dữ liệu Oxford 5K+100K.

Kết quả thí nghiệm trên cả ba bộ dữ liệu đều cho thấy hiệu quả của việc tích hợp thông tin không gian ảnh vào chỉ mục ngược. Các thí nghiệm trên đều sử dụng thông số  $L = 2$  ( $L$  là thông số để thiếp lập cho cấp cao nhất của không gian phân cấp). Để kiểm tra sự phụ thuộc của kết quả vào  $L$ , chúng tôi cũng đã đo đặc kết quả với các mức  $L$  khác nhau. Chi tiết được thể hiện trong [Bảng 4.7](#) và [Bảng 4.8](#) cùng các biểu đồ trong [Hình 4.5](#). Có thể thấy rõ rằng khi giá trị của  $L$  tăng thì độ chính xác cũng tăng. Điều đó nghĩa là khi tích hợp thông tin không gian ảnh với những lưới ô vuông phân cấp càng dày thì sự khác nhau giữa các hình ảnh sẽ càng được thể hiện rõ nét hơn. Tuy nhiên, không phải lúc nào  $L$

## 4. Thực nghiệm và đánh giá kết quả



Hình 4.4: Biểu đồ so sánh độ chính xác và thời gian truy vấn giữa các phương pháp trên bộ Oxford 5K+100K.

$L$	mAP	Số tập chỉ mục ngược	Thời gian truy vấn (55 truy vấn)	Bộ nhớ sử dụng
$L = 0$	0.5678	1	0.0794 (s)	68.31MB
$L = 1$	0.5791	5	0.1092 (s)	183.17MB
$L = 2$	<b>0.5851</b>	<b>21</b>	<b>0.1651 (s)</b>	<b>418.69MB</b>
$L = 3$	0.5779	85	0.1806 (s)	1.48GB

Bảng 4.7: Hiệu suất của phương pháp đề xuất với các giá trị khác nhau của  $L$  trên bộ dữ liệu Oxford 5K.

tăng thì độ chính xác cũng sẽ tăng theo.

Hình 4.6 cho thấy ví dụ về hình ảnh truy vấn và thể hiện các kết quả trả về với các phương pháp khác nhau trên bộ Oxford 5K. 10 hình ảnh có đúng đầu trong kết quả trả về của các phương pháp được hiển thị. Có thể thấy rằng phương pháp đề xuất của chúng tôi có độ chính xác tương đương với phương pháp cơ sở 2 trong khi kết quả của phương pháp cơ sở 1 vẫn chứa một vài hình ảnh sai.

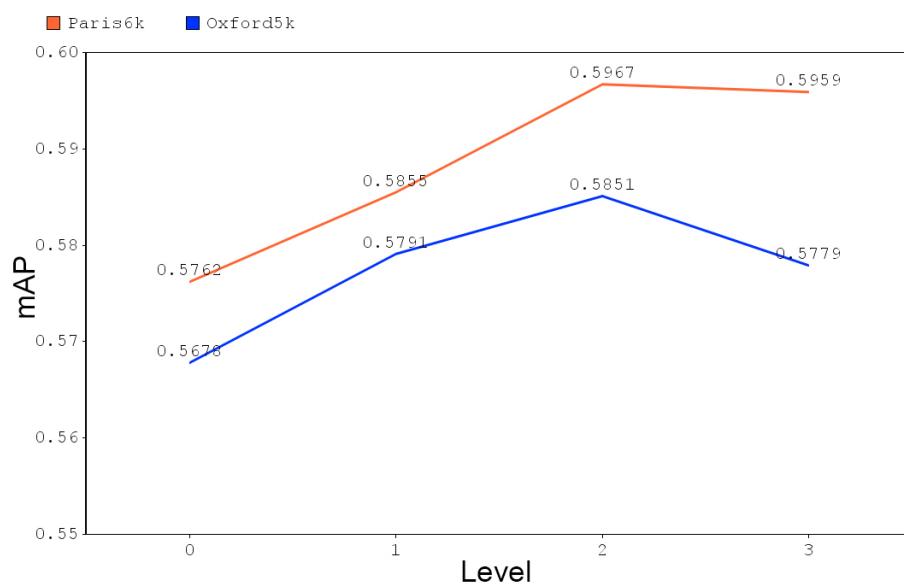
Tương tự, trong hình 4.7, phương pháp đề xuất cũng cho thấy độ chính xác tốt tương đương với phương pháp cơ sở 2 khi chạy trên bộ dữ liệu Paris 6K.

#### 4. Thực nghiệm và đánh giá kết quả

---

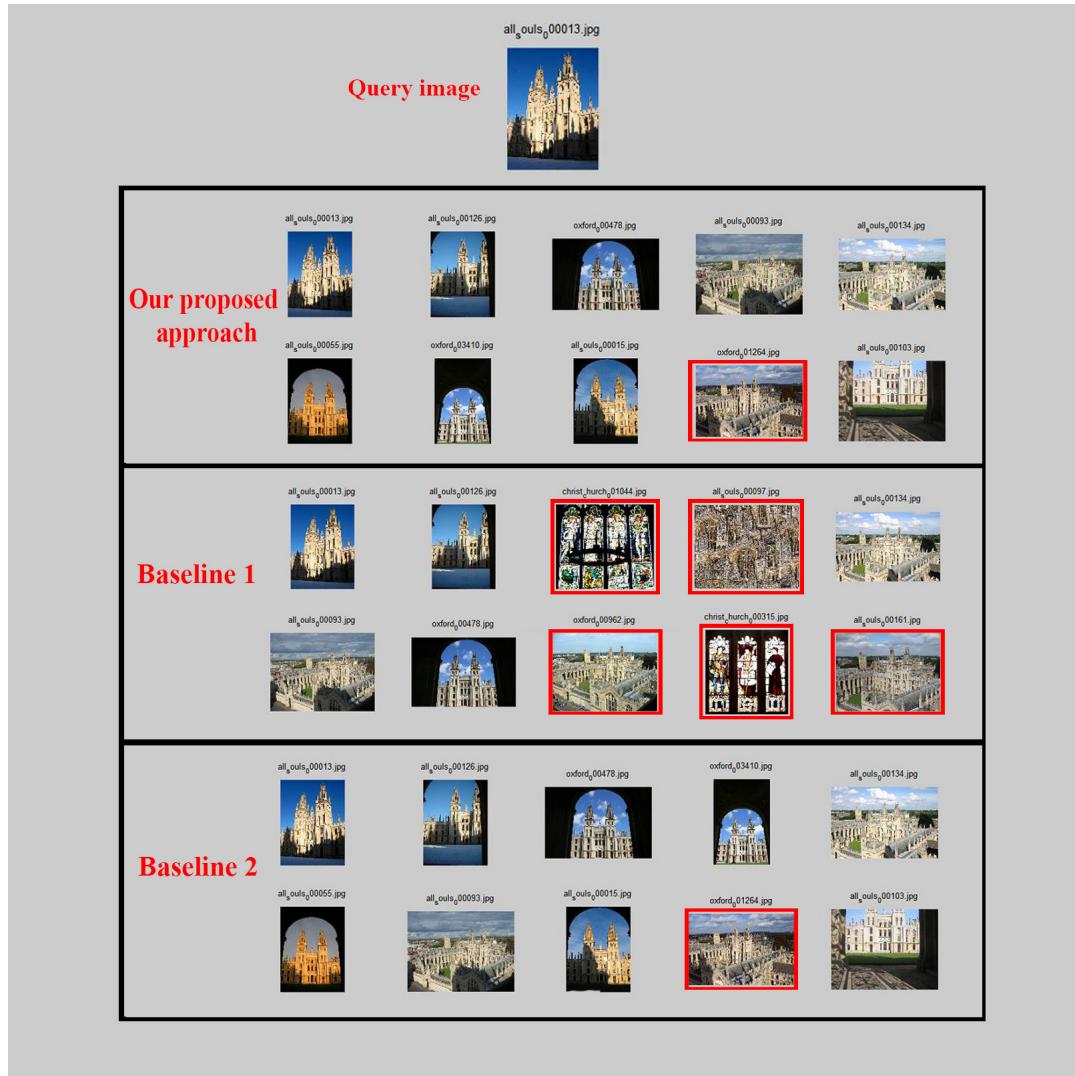
$L$	mAP	Số tập chỉ mục ngược	Thời gian truy vấn (55 truy vấn)	Bộ nhớ sử dụng
$L = 0$	0.5762	1	0.1138 (s)	80.37MB
$L = 1$	0.5855	5	0.1523 (s)	207.68MB
$L = 2$	<b>0.5967</b>	<b>21</b>	<b>0.2158 (s)</b>	<b>519.01MB</b>
$L = 3$	0.5959	85	0.2953 (s)	1.53GB

Bảng 4.8: Hiệu suất của phương pháp đề xuất với các giá trị khác nhau của  $L$  trên bộ dữ liệu Paris 6K.



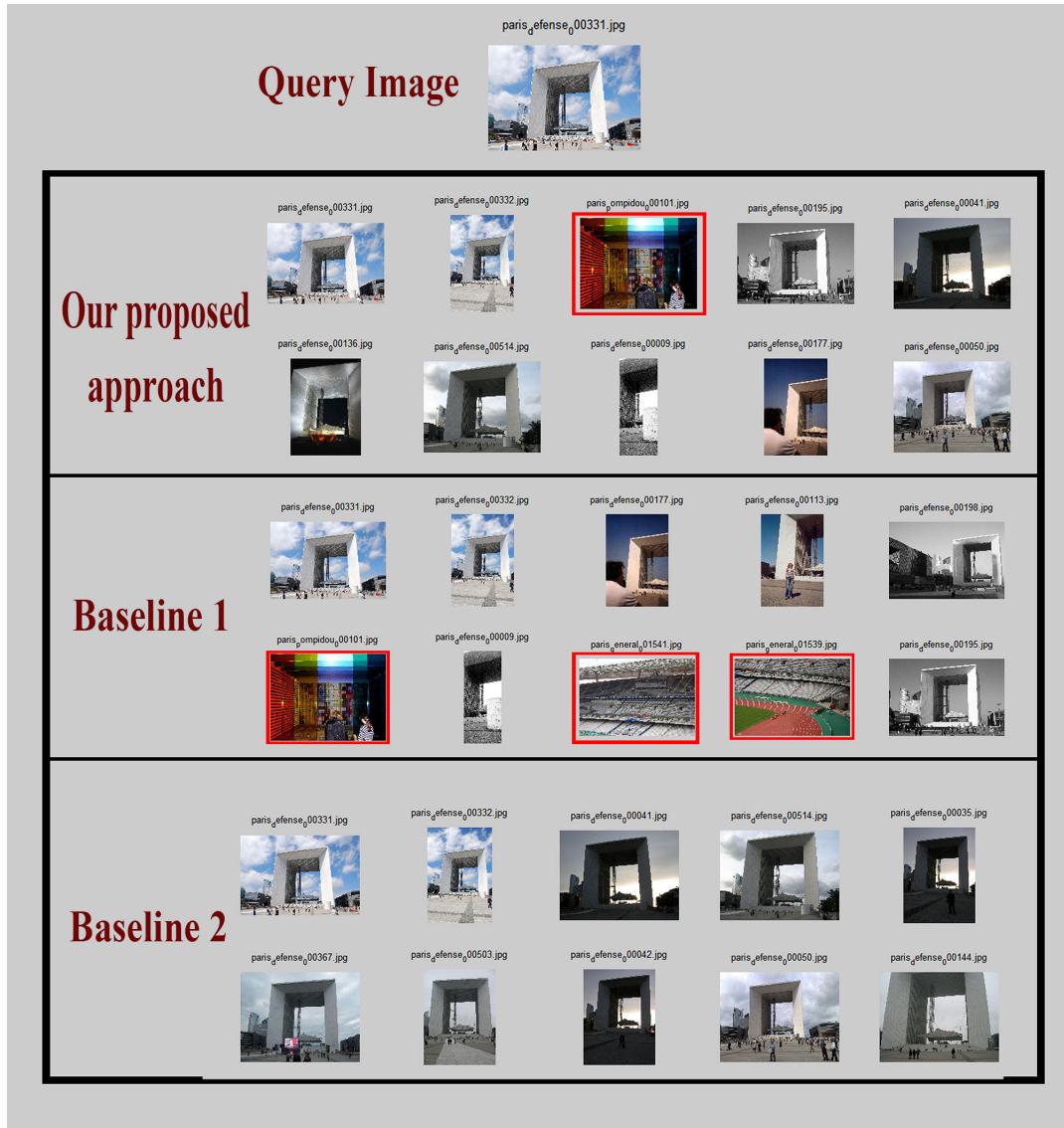
Hình 4.5: Biểu đồ hiệu suất của phương pháp đề xuất trên các cấp độ phân cấp  $L$  khác nhau.

#### 4. Thực nghiệm và đánh giá kết quả



Hình 4.6: Ví dụ thể hiện kết quả khi tìm kiếm một hình ảnh trên bộ dữ liệu Oxford 5K. Những kết quả sai được đánh dấu bằng ô có viền màu đỏ.

#### 4. Thực nghiệm và đánh giá kết quả



Hình 4.7: Ví dụ thể hiện kết quả khi tìm kiếm một hình ảnh trên bộ dữ liệu Paris 6K. Những kết quả sai được đánh dấu bằng ô có viền màu đỏ.

# Chương 5

## Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh

Trên cơ sở các phương pháp đã nghiên cứu và những kết quả thực nghiệm của phương pháp đề xuất, chúng tôi tiến hành xây dựng ứng dụng tìm kiếm đối tượng trên ảnh nhằm thực nghiệm phương pháp đề xuất trên môi trường thực tế chứng minh tính thực tiễn của đề tài.

Trong chương này, trước tiên chúng tôi sẽ giới thiệu tổng quan về mục đích và các chức năng chính của ứng dụng (mục 5.1). Sau đó là bước thiết kế kiến trúc, tổ chức các thành phần và giao diện của ứng dụng (mục 5.2). Cuối cùng là bước cài đặt, thử nghiệm và đánh giá kết quả của hệ thống đã cài đặt (mục 5.3).

### 5.1 Tổng quan ứng dụng

#### 5.1.1 Mục đích và phạm vi của ứng dụng

Như đã giới thiệu trong mục 1.1, các hệ thống truy vấn ảnh có vô vàn ứng dụng khác nhau trong thực tế. Trong đề tài này, chúng tôi xây dựng ứng dụng nhằm phục vụ mục đích cơ bản là tìm kiếm đối tượng trên những kho dữ liệu ảnh với kích thước có thể lên tới hàng trăm ngàn ảnh. Đó có thể là kho dữ liệu ảnh của một tổ chức, công ty về một lĩnh vực nào đó. Hay xa hơn, ứng dụng này có thể phát triển cho việc tìm kiếm đối tượng trên kho dữ liệu video vì ta hoàn toàn có thể rút trích được hình ảnh từ các frame của video.

## **5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh**

---

Hơn nữa, ngày nay điện thoại thông minh đang dần trở lên rất phổ biến và dần trở thành vật bất ly thân của con người. Các ứng dụng trên điện thoại thông minh cũng phát triển ồ ạt, cung cấp cho người dùng rất nhiều tiện ích. Vì vậy chúng tôi hướng tới một ứng dụng giao tiếp với người dùng trên nền tảng điện thoại thông minh sử dụng hệ điều hành Android. Và để gần hơn với các yêu cầu thực tế khi số lượng người dùng ngày càng tăng, chúng tôi đề xuất xây dựng một hệ thống xử lý đám mây với nhiều server xử lý, kết nối với nhau thông qua dịch vụ web (web service).

### **5.1.2 Các chức năng chính**

Ứng dụng được xây dựng với mục tiêu cung cấp một hệ thống tìm kiếm đối tượng trên ảnh, nhận đầu vào là hình ảnh chứa đối tượng truy vấn và trả về danh sách xếp hạng các hình ảnh theo mức độ tương đồng với hình truy vấn. Chi tiết các chức năng chính của ứng dụng như sau:

- Chụp hình đối tượng: người dùng có thể sử dụng camera của điện thoại để chụp hình đối tượng và tìm kiếm.
- Chọn ảnh được lưu trữ trước trong máy: người dùng có thể chọn một ảnh có chứa đối tượng được lưu trữ trong điện thoại, thay nhớ để tìm kiếm.
- Chọn vùng đối tượng: để nâng cao độ chính xác của việc tìm kiếm, người dùng có thể khoanh vùng đối tượng cần tìm trên ảnh.
- Xem và tương tác với kết quả truy vấn: hiển thị kết quả truy vấn là danh sách các hình ảnh và cho phép người dùng xem và tải xuống các hình ảnh này.

## **5.2 Xây dựng ứng dụng**

### **5.2.1 Kiến trúc tổng quan**

Kiến trúc của hệ thống gồm làm 3 phần chính:

- Client side: là một ứng dụng trên điện thoại di động sử dụng hệ điều hành Android, cung cấp giao diện để người dùng tương tác trực tiếp với hệ thống tìm kiếm đối tượng trên ảnh.
- Web service: là thành phần trung gian, nhận các yêu cầu từ phía client.

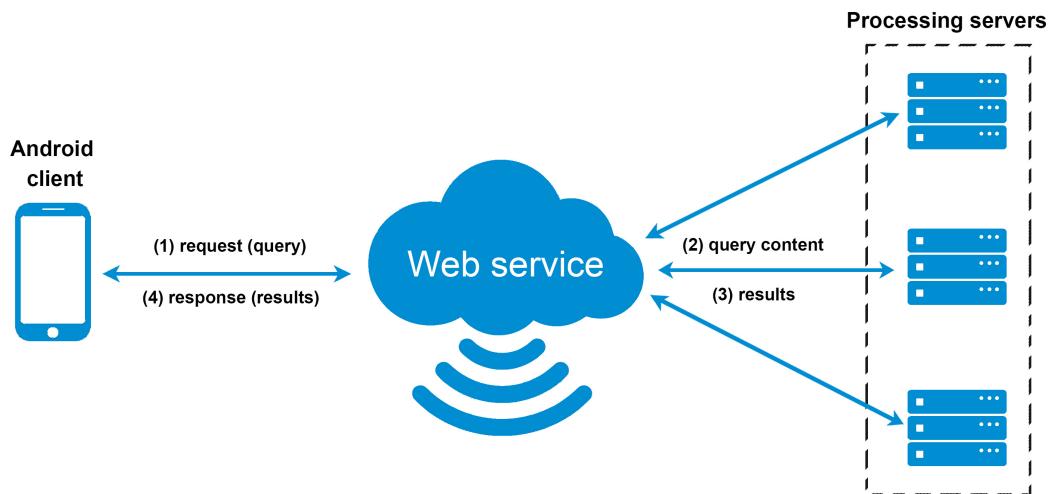
## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh

Thực hiện điều phối và cân bằng tải, chuyển tiếp các yêu cầu từ người dùng tới các server xử lý, đồng thời tiếp nhận kết quả xử lý và trả về cho từng client tương ứng.

– Server side: là thành phần xử lý chính của hệ thống, nhận yêu cầu từ web service chuyển tiếp tới, xử lý và trả về kết quả cho web service. Chi phí xử lý truy vấn hình ảnh tại server rất lớn, do đó để đảm bảo yêu cầu thực tế, hệ thống có thể sử dụng nhiều server để xử lý đồng thời các yêu cầu.

Hình 5.1 minh họa cho mô hình hoạt động tổng quan của hệ thống.

Chi tiết các thành phần được trình bày trong các phần tiếp theo (??, 5.2.3 5.2.4).



Hình 5.1: Kiến trúc tổng quan của hệ thống.

## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh

---

### 5.2.2 Server side

Server side chính là thành phần quan trọng nhất của hệ thống, thực hiện các tác vụ xử lý truy vấn. Cấu trúc của server side gồm hai thành phần chính tương ứng với hai quá trình xử lý là quá trình training và quá trình truy vấn.

– Quá trình training: là quá trình biểu diễn lại tập dữ liệu hình ảnh, sử dụng các mô hình biểu diễn hình ảnh kết hợp với các phương pháp, kỹ thuật khác với mục đích tạo ra các dữ liệu cần thiết phục vụ quá trình truy vấn.

– Quá trình truy vấn: là quá trình tương tác trực tiếp với hệ thống tìm kiếm đối tượng trên ảnh, nhận đầu vào là một hình ảnh, thực hiện các bước xử lý truy vấn ảnh và trả về kết quả là một tập các hình ảnh được sắp xếp theo thứ tự có độ tương đồng với hình ảnh truy vấn giảm dần.

#### 5.2.2.1 Quá trình training

Như đã trình bày sơ lược trong mục trong mục [3.1](#) và [4.2](#), quá trình training bao gồm các bước sau:

**Dò tìm, phát hiện và rút trích đặc trưng.** Từ kho dữ liệu hình ảnh được lưu sẵn trên server, bước đầu tiên của quá trình xử lý là sử dụng phương pháp phát hiện keypoint Hessian-Affine để xác định được vị trí các điểm keypoint trên hình ảnh. Từ thông tin đó, bộ SIFT descriptor mô tả và rút trích được 1 vector 128 chiều tương ứng với mỗi điểm keypoint. Những vector này sẽ được tính toán lại bằng phương pháp RootSIFT[[43](#)]. Các vector thu được chính là các tập đặc trưng của hình ảnh.

**Gom cụm đặc trưng và xây dựng từ điển.** Các vector đặc trưng sẽ được gom cụm bằng thuật toán Approximate K-Means[[43](#)] để thu được các visual word với số lượng cụm là  $k = 1$  triệu. Kết quả ta sẽ thu được các visual word để mô tả hình ảnh. Từ những visual word này, ta sẽ xây dựng thành một từ điển để phục vụ cho việc biểu diễn hình ảnh khi truy vấn.

**Xây dựng chỉ mục ngược.** Theo như phương pháp đề xuất đã được đề cập chi tiết trong [Chương 3](#), chúng tôi sẽ xây dựng chỉ mục ngược từ từ điển nhằm đạt được hiệu suất cao trong quá trình truy vấn. Cụ thể hơn, từ bộ từ điển và thông tin không gian của tất cả các visual word của mỗi hình ảnh trong bộ dữ liệu, với cấp độ phân cấp là 2, có 21 chỉ mục ngược sẽ được tạo ra và lưu vào một

## **5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh**

---

file.

Sơ đồ xử lý cùng các thông số cài đặt được trình bày chi tiết trong phần cài đặt thí nghiệm (Mục 4.2) và trong Hình 3.1. Quá trình training được thực hiện độc lập trên server và là bước tiền đề, tạo dữ liệu phục vụ cho quá trình truy vấn.

### **5.2.2.2 Quá trình truy vấn**

Đây là quá trình tương tác trực tiếp với hệ thống tìm kiếm và chạy song song web service, nhận yêu cầu từ web service chuyển tiếp tới sau đó tiến hành truy vấn trả kết quả ngược trở lại web service.

Ngày nay, chất lượng hình ảnh ngày càng được tăng cao, kéo theo dung lượng hình ảnh cũng rất lớn. Việc trả về cho client một lượng lớn hình ảnh chất lượng cao là không thể vì lưu lượng và tốc độ băng thông còn bị giới hạn rất nhiều. Vì vậy sau khi xử lý truy vấn, server sẽ trả về cho client một danh sách hình ảnh thu nhỏ để đảm bảo tốc độ phản hồi. Chi tiết sẽ được trình bày rõ hơn trong phần tối ưu hiệu suất hệ thống (Mục 5.2.5).

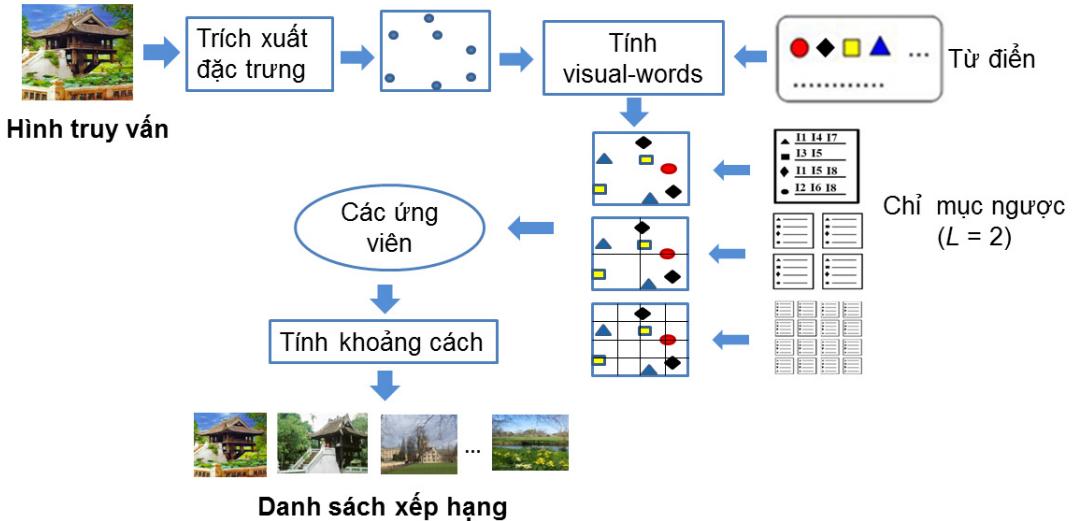
Việc trả về cho client các hình ảnh thu nhỏ sẽ phát sinh thêm các chức năng khác như: xem hình ảnh chất lượng tốt hơn, tải hình ảnh gốc về thiết bị. Vì vậy, tại phía server cung cấp hai chức năng chính tương ứng với hai loại yêu cầu từ phía client là yêu cầu truy vấn một hình ảnh và yêu cầu tải một (hoặc một tập) các hình ảnh theo các mức chất lượng khác nhau.

Với yêu cầu như trên, các server được chia làm 2 loại chính:

- Server truy vấn: tiếp nhận yêu cầu truy vấn hình ảnh, xử lý và trả về kết quả là danh sách xếp hạng cùng một phần các hình ảnh thu nhỏ giống với hình truy vấn nhất.
- Server truy xuất dữ liệu: tiếp nhận yêu cầu truy xuất hình ảnh, trả về hình ảnh được lấy từ bộ nhớ.

Quá trình xử lý truy vấn được đã trình bày chi tiết trong mục 3.2 và được thể hiện trong Hình 5.2. Đầu tiên, ta sẽ rút trích đặc trưng từ hình ảnh truy vấn. Sau đó các đặc trưng này được đưa vào từ điển để lấy được các visual word tương ứng. Từ các visual word và vị trí của nó trên hình ảnh, ta sẽ dùng chỉ mục ngược để xuất để tìm kiếm các hình ảnh ứng viên có liên quan và đồng thời tính

## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh



Hình 5.2: Quá trình truy vấn tại server.

khoảng cách từ hình truy vấn tới các hình ảnh này. Sau đó, danh sách hình ảnh ứng viên sẽ được sắp xếp theo mức độ tương tự từ cao tới thấp. Kết quả sẽ được server gửi ngược trở về webservice và trả về cho client yêu cầu.

### 5.2.3 Web service

Web service là thành phần trung gian đóng vai trò phân phối các yêu cầu từ phía client tới các server xử lý. Thành phần này sử dụng dịch vụ web chạy trên các máy chủ web cung cấp các phương thức giúp client và server xử lý kết nối, truyền tải các dữ liệu cần thiết. Có nhiều cách để kết nối ba thành phần này với nhau, chi tiết sẽ được trình bày dưới đây.

Với một hệ thống tìm kiếm cơ bản sử dụng từ khóa thông thường, thành phần xử lý có thể chính là web service. Với mỗi yêu cầu tìm kiếm, nó sẽ tạo ra một tiến trình và truy xuất tới cơ sở dữ liệu để lấy thông tin một cách nhanh chóng, có thể đáp ứng nhiều yêu cầu đồng thời, phụ thuộc vào phần cứng và các thiết lập trên nó. Nhưng với một hệ thống tìm kiếm hình ảnh, chi phí tính toán và tài nguyên phục vụ quá trình tìm kiếm rất lớn, việc tạo ra nhiều tiến trình xử lý là không khả thi. Nếu giới hạn số lượng yêu cầu ở mức thấp để đảm bảo an toàn

## **5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh**

---

cho hệ thống thì sẽ không đáp ứng được yêu cầu thực tế, ngược lại nếu không kiểm soát các yêu cầu truy vấn thì hệ thống có thể sẽ không đáp ứng được và dễ dàng bị quá tải dẫn đến nguy hiểm cả cho phần cứng. Chính vì vậy, giải pháp tốt nhất chính là chia tải, bằng cách sử dụng nhiều server xử lý và chuyển tiếp các yêu cầu truy vấn tới các server này. Từ đó ta có hai giải pháp kết nối:

– Giải pháp 1: Web service lưu giữ thông tin các server xử lý, khi có yêu cầu (request) từ phía client, nó sẽ gửi thông tin server để client trực tiếp kết nối và truyền tải dữ liệu tới server xử lý.

– Giải pháp 2: Web service nhận yêu cầu từ phía client, truyền tải dữ liệu tới server xử lý, chờ kết quả xử lý từ server và gửi lại cho client. Như vậy client không được phép kết nối trực tiếp tới server xử lý.

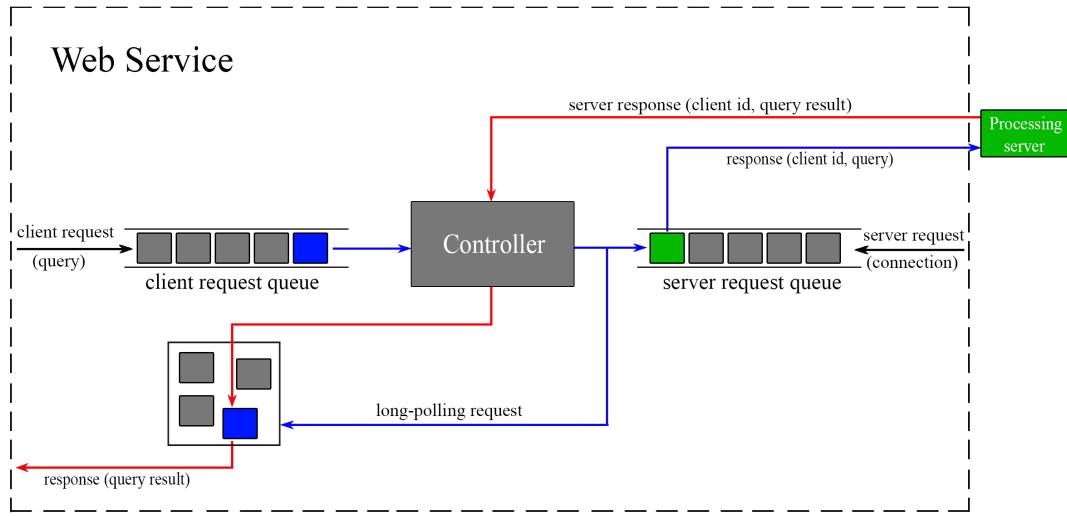
Có thể thấy với giải pháp thứ nhất, dữ liệu qua lại trực tiếp giữa client và server nên tối ưu được quá trình truyền dẫn, không cần qua trung gian. Nhưng do client có thể nắm được thông tin của server nên nó gấp phải một vấn đề lớn là bảo mật. Vì vậy chúng tôi không sử dụng giải pháp này.

Với giải pháp thứ 2, client chỉ được phép kết nối tới web service và không hề biết thông tin của bất cứ server xử lý nào nên vấn đề bảo mật đã được giải quyết. Nhược điểm của cách này là dữ liệu phải thông qua trung gian, làm tăng thời gian truyền tải. Nhưng thường các server kết nối với nhau qua một mạng lưới internet hoặc mạng lan với tốc độ băng thông rất lớn nên thời gian truyền tải dữ liệu phần lớn phụ thuộc vào tốc độ băng thông của client. Vậy giải pháp này rõ ràng tốt hơn giải pháp đầu tiên.

Có rất nhiều cách để xây dựng một hệ thống với giải pháp trên. Một trong những cách đơn giản mà chúng tôi sử dụng là dùng web service cung cấp các phương thức để truyền tải dữ liệu giữa client và server xử lý. Cụ thể hơn, nó giống như một hệ thống chat giữa các user với nhau và web service là thành phần chuyển tiếp các thông điệp. Khi đó client và server sẽ "chat" với nhau, nhưng thông điệp ở đây phức tạp hơn. Chẳng hạn khi client truy vấn một hình ảnh, thì thông điệp ở đây là yêu cầu truy vấn kèm theo hình ảnh đã được mã hóa, và khi hoàn tất xử lý thì server sẽ gửi lại thông điệp là danh sách xếp hạng cùng cách hình ảnh.

Sơ đồ mô hình hoạt động chi tiết của web service được thể hiện trong [Hình 5.3](#).

## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh



Hình 5.3: Sơ đồ mô hình hoạt động chi tiết của web service.

Trước tiên, mỗi yêu cầu do client gửi lên sẽ được đưa vào một hàng đợi. Đồng thời kết nối từ các server gửi lên cũng được đưa vào một hàng đợi để chờ client. Thành phần xử lý trung tâm là Controller sẽ lấy từng yêu cầu của client trả về cho một server đang chờ kết nối. Tiếp đó, yêu cầu của client sẽ bị giữ lại để chờ kết quả xử lý từ phía server bằng kỹ thuật long-polling. Sau khi hoàn tất xử lý, server sẽ gửi lại web service kết quả xử lý cùng với thông tin client, từ đó web service sẽ đáp trả yêu cầu của client tương ứng.

**Kỹ thuật Long-polling.** Để client và server xử lý kết nối thời gian thực với nhau, chúng tôi sử dụng kỹ thuật long-polling là một kỹ thuật trong lập trình ứng dụng web. Một ví dụ đơn giản về kỹ thuật này là ứng dụng chat trên internet. Khi một user (1) gửi thông điệp chat tới một user khác (2) thông qua web service, nhưng web service không thể gửi thông điệp đó tới user 2. Khi đó user 2 phải liên tục gửi yêu cầu lên web service để xem có ai chat với mình hay không. Như vậy nếu muốn kết nối thời gian thực thì các user phải liên tục gửi yêu cầu lên web service, với số lượng lớn yêu cầu thì web service sẽ không thể đáp ứng nổi. Khi đó kỹ thuật long-polling được đưa ra nhằm giải quyết vấn đề này. Trong trường hợp trên, khi user 1 gửi yêu cầu lên web service, yêu cầu này sẽ không được trả về ngay lập tức mà nó bị giữ lại. Khi user 2 gửi thông điệp tới

## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh

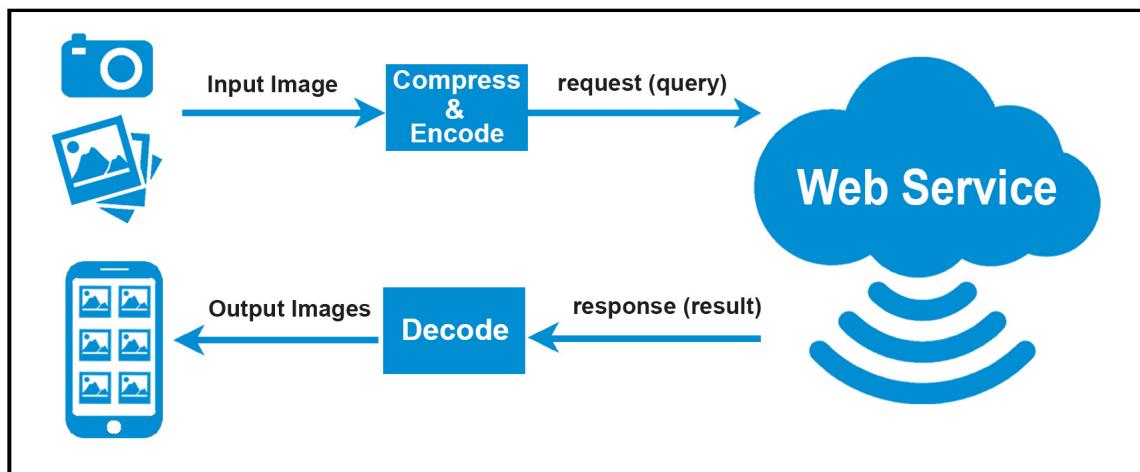
user 1 thì lúc này web service mới trả yêu cầu cùng thông điệp cho user 1. Cứ như thế 2 user có thể tương tác thời gian thực với nhau.

Trong hệ thống của chúng tôi, kỹ thuật long-polling được sử dụng để giữ yêu cầu kết nối của các server để chờ client, và giữ yêu cầu truy vấn của client để chờ kết quả xử lý từ server.

### 5.2.4 Client side

#### 5.2.4.1 Tổng quan

Client side là một ứng dụng cung cấp giao diện để người dùng tương tác với hệ thống tìm kiếm. Nhận đầu vào là một hình ảnh cùng vùng đối tượng cần truy vấn, ứng dụng sẽ nén và mã hóa hình ảnh này sau đó gửi tới web service thông qua giao thức HTTP. Khi có kết quả trả về, ứng dụng sẽ phân tích và giải mã kết quả sau đó hiển thị cho người dùng. Hình 5.4 sau đây cho mô hình hoạt động tổng quan của ứng dụng tại client side.

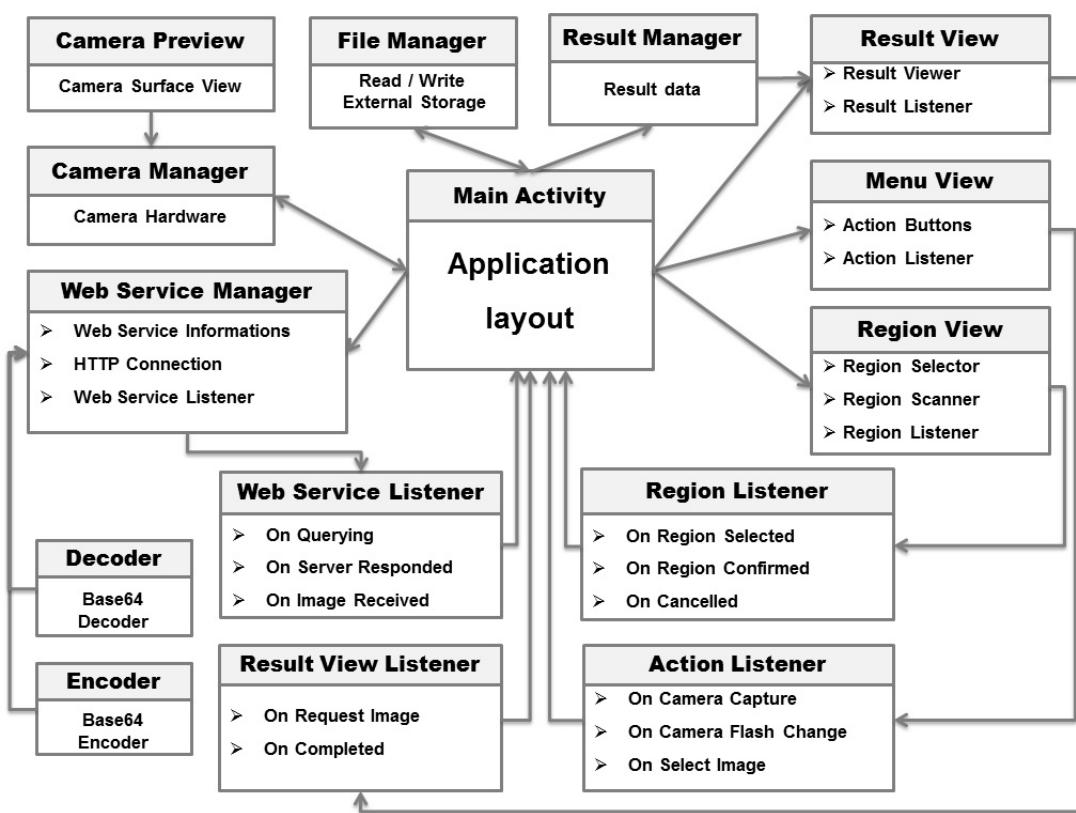


Hình 5.4: Kiến trúc của client side.

Ứng dụng được xây dựng dành cho hệ điều hành Android, được viết bằng ngôn ngữ lập trình Java. Chi tiết tổ chức các lớp được thể hiện trong Hình 5.5.

- **Lớp CameraManager:** truy xuất tới camera trên điện thoại di động, thiết lập các thông số liên quan tới camera, nhận dữ liệu hình ảnh từ camera.

## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh



Hình 5.5: Sơ đồ tổ chức các lớp của ứng dụng trên hệ điều hành Android.

## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh

---

- **Lớp WebServiceManager:** lưu trữ thông tin kết nối và thực hiện các kết nối tới web service, gửi và nhận dữ liệu truy vấn, gọi bộ interface lắng nghe sự kiện WebServiceListener để tương tác với các thành phần khác của ứng dụng.
- **Lớp ResultManager:** nhận, lưu trữ và quản lý kết quả truy vấn.
- **Lớp Decoder và Encoder:** giải mã và mã hóa hình ảnh. Để dễ dàng cho quá trình truyền dẫn dữ liệu giữa client, web service và server thì mỗi hình ảnh sẽ được mã hóa dưới dạng Base64 String.
- **Interface WebServiceListener:** bộ lắng nghe các sự kiện liên quan tới web service. OnQuerying: trong khi đang truy vấn. OnServerResponded: khi có kết quả phản hồi từ server, OnImageReceived: khi có kết quả là các hình ảnh được trả về.
- **Interface ResultViewListener:** bộ lắng nghe sự kiện liên quan tới việc hiển thị kết quả truy vấn. OnRequestImage: yêu cầu tải về một hình ảnh khi chọn xem hoặc tải xuống thiết bị, tải về danh sách hình ảnh khi người dùng muốn xem thêm kết quả. OnCompleted: khi hoàn tất việc xem kết quả, kết thúc một quá trình truy vấn.
- **Interface ActionListener:** bộ lắng nghe các sự kiện là các chức năng trên ứng dụng. OnCameraCapture: khi có yêu cầu chụp hình. OnCameraFlashChange: khi có yêu cầu thay đổi trạng thái đèn flash. OnSelectImage: khi có yêu cầu chọn một hình ảnh trong thiết bị để truy vấn.
- **Interface RegionListener:** bộ lắng nghe các sự kiện liên quan tới chức năng chọn vùng đối tượng trên camera. OnRegionSelected: khi người dùng đã chọn vùng đối tượng. OnRegionConfirmed: khi vùng đối tượng đã được xác nhận và được tiến hành truy vấn. OnRegionCancelled: Khi vùng đối tượng đã bị hủy.
- **Lớp CameraPreview:** hiển thị dữ liệu được lấy từ camera.
- **Lớp MenuView:** cung cấp các phím chức năng cho phép người dùng thao tác với ứng dụng, bao gồm phím chụp hình cho phép người dùng chụp hình

## **5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh**

---

ảnh để truy vấn, phím flash để người dùng thay đổi trạng thái đèn flash, phím chọn hình ảnh để người dùng chọn một hình ảnh trong thiết bị dùng làm hình ảnh truy vấn; gọi bộ lắng nghe sự kiện ActionListener khi người dùng thực hiện các chức năng để tương tác với các thành phần khác.

- **RegionView:** cung cấp cho người dùng chức năng chọn vùng đối tượng trên camera. gọi bộ lắng ghe sự kiện RegionSelectionListener để tương tác với các thành phần khác.
- **ResultView:** cung cấp giao diện cho phép người dùng xem và thao tác trên kết quả truy vấn, gọi bộ lắng nghe sự kiện ResultListener để thao tác với các thành phần khác.
- **Lớp FileManager:** đọc, ghi dữ liệu từ bộ nhớ thiết bị.
- **Lớp MainActitity:** cung cấp giao diện chính cho ứng dụng, là thành phần trung tâm quản lý toàn bộ các lớp khác, được cài đặt tất cả các bộ lắng nghe sự kiện để các lớp có thể giao tiếp với nhau thông qua nó.

### **5.2.4.2 Chức năng và giao diện**

Là thành phần tương tác trực tiếp với người dùng, do đó ứng dụng cần đáp ứng các yêu cầu về giao diện thân thiện, đơn giản, tiện dụng nhưng vẫn cung cấp đầy đủ các chức năng cho người dùng.

Dưới đây là giao diện cùng các chức năng của ứng dụng được thể hiện qua các hình [5.6](#), [5.7](#), [5.8](#), [5.9](#), [5.10](#) và [5.11](#).

### **5.2.5 Tối ưu hiệu suất hệ thống**

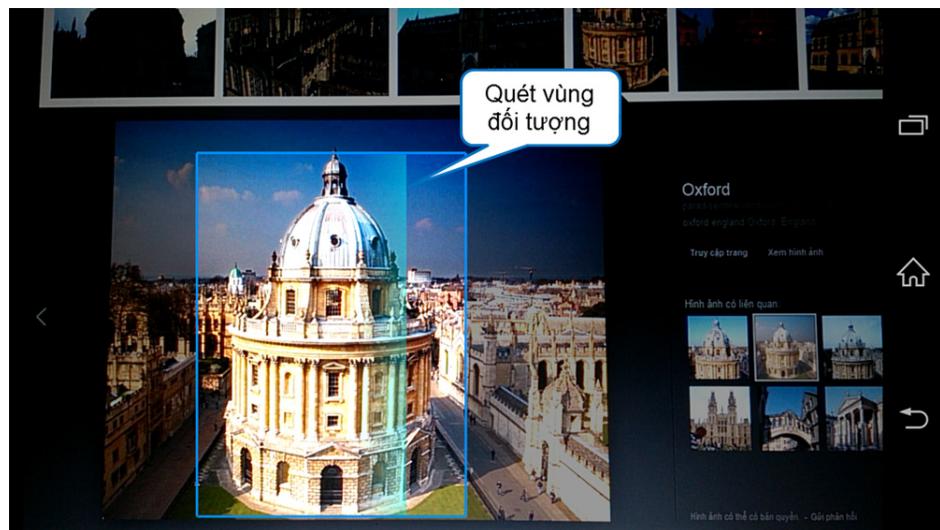
Việc xử lý truy vấn hình ảnh đòi hỏi chi phí tính toán rất lớn, vì vậy một hệ thống tìm kiếm hình ảnh muốn đáp ứng được các yêu cầu thực tế cần phải tối ưu hóa hiệu suất trong quá trình xử lý cũng như truyền dẫn. Trong hệ thống đã xây dựng, chúng tôi đưa ra các giải pháp để tối ưu hoạt động tại cả ba thành phần của hệ thống là client side, web service và sever side.

- **Tại client side: giảm tối đa dung lượng hình ảnh truy vấn**

## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh

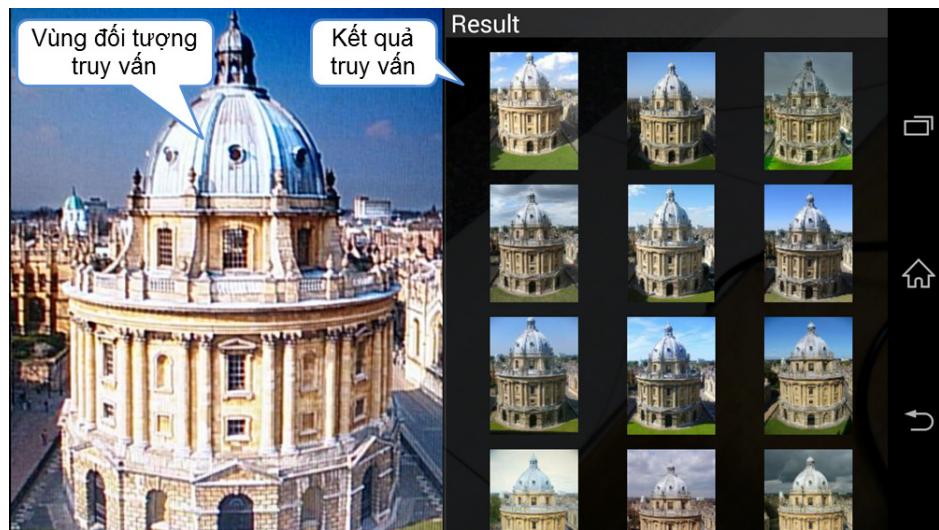


Hình 5.6: Hình ảnh giao diện ứng dụng với các chức năng chính. Ứng dụng cho phép người dùng chọn vùng đối tượng trên camera, chụp hình vùng đối tượng, bật / tắt đèn flash và chọn hình ảnh trong thiết bị để truy vấn.

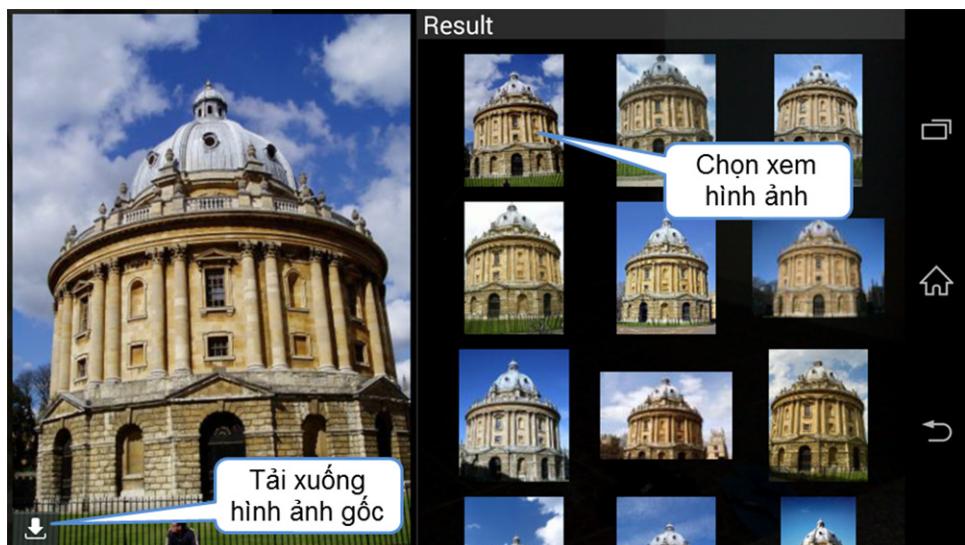


Hình 5.7: Hình ảnh ứng dụng trong khi thực hiện truy vấn. Trong khi chờ kết quả xử lý từ server, ứng dụng sẽ tạo hiệu ứng quét trên vùng đối tượng đã chọn để tránh cảm giác chờ đợi. Tương tự khi chọn một hình ảnh trong thiết bị để truy vấn, ứng dụng cũng quét trên toàn bộ vùng truy vấn.

## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh

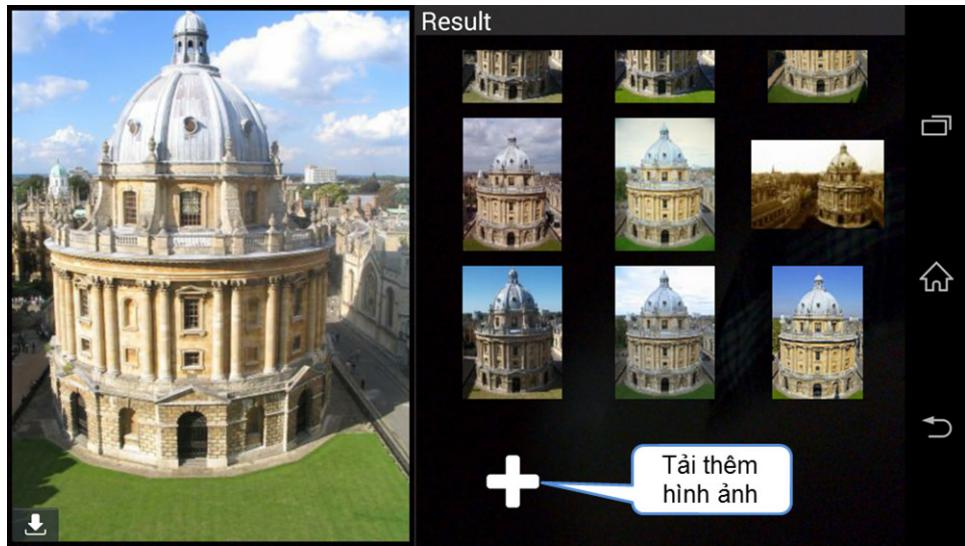


Hình 5.8: Hình ảnh ứng dụng khi có kết quả trả về. Ứng dụng cho người dùng xem vùng đối tượng vừa được truy vấn cùng danh sách các hình ảnh thu nhỏ được trả về từ server.

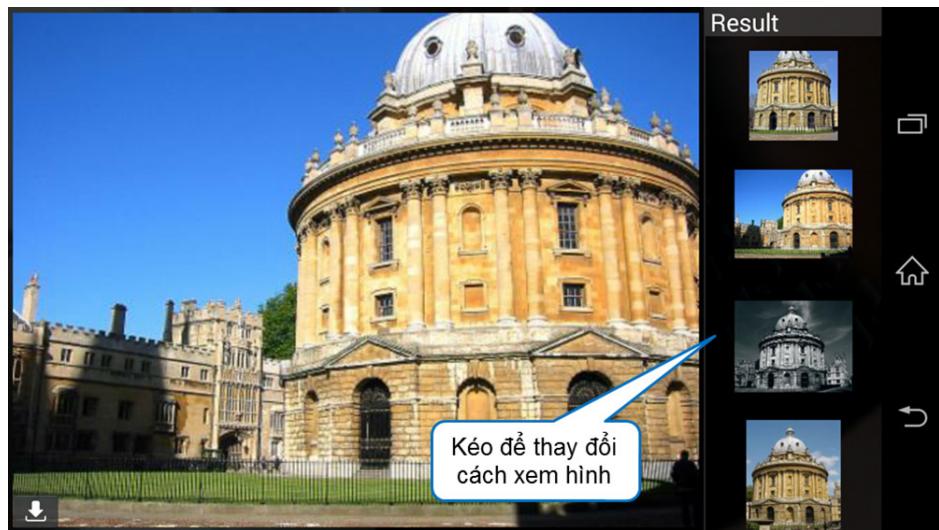


Hình 5.9: Hình ảnh ứng dụng khi xem chi tiết kết quả truy vấn. Người dùng có thể chọn xem một hình ảnh trong danh sách kết quả truy vấn. Ứng dụng sẽ tải hình ảnh chất lượng cao hơn để người dùng dễ dàng xem trong khung hình lớn hơn, nhưng đây chưa phải hình ảnh với chất lượng tốt nhất. Do đó, ứng dụng cũng cho phép người dùng tải hình ảnh gốc chất lượng cao bằng cách chạm vào nút như trên hình.

## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh



Hình 5.10: Hình ảnh ứng dụng với chức năng tải thêm hình ảnh trong danh sách kết quả truy vấn. Khi hoàn tất xử lý, server chỉ trả về cho ứng dụng một số hình ảnh đúng đầu danh sách xếp hạng, người dùng có thể tải thêm các hình ảnh khác bằng cách chạm vào dấu cộng như trên hình, với số lượng hình ảnh tối đa có thể tải là 50 hình.



Hình 5.11: Ứng dụng cho phép người dùng thay đổi chế độ xem hình ảnh truy vấn bằng cách kéo khung kết quả để thu nhỏ, phóng to hình ảnh đang xem.

## **5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh**

---

Như đã trình bày trong mục [5.2.3](#), thời gian truyền dữ liệu dẫn phụ thuộc phần lớn vào tốc độ băng thông tại client, mà hình ảnh là dữ liệu khá lớn. Vì vậy chúng tôi đưa ra giải pháp giảm tối đa dung lượng hình ảnh nhưng vẫn đảm bảo kết quả truy vấn.

Dung lượng hình ảnh phụ thuộc vào ba yếu tố chính là kích thước ảnh, định dạng ảnh và chất lượng hình ảnh. Khi nhận được dữ liệu từ camera của điện thoại, hình ảnh sẽ được nén dưới định dạng JPEG ở mức chất lượng là 30%, sau đó nó sẽ được điều chỉnh lại với kích thước tối đa là 500 x 500 pixel. Ví dụ, một hình ảnh được chụp từ điện thoại với độ phân giải 2 Megapixel (1920x1080) có dung lượng khoảng 500 kilobytes (kB), sau khi được nén dưới định dạng JPEG ở mức chất lượng 30% và điều chỉnh kích thước xuống 500x281 pixel, dung lượng hình ảnh giảm chỉ còn khoảng 15kB, tức là giảm 97%. Như vậy, việc truyền tải sẽ rất nhanh chóng, có thể đáp ứng được trên mạng EDGE với tốc độ 236kb/s. Nhưng với chất lượng hình ảnh như vậy sẽ ảnh hưởng tới kết quả truy vấn.

Kết quả truy vấn phụ thuộc phần lớn vào quá trình trích xuất đặc trưng tại server xử lý, và chất lượng hình ảnh sẽ ảnh hưởng trực tiếp tới quá trình này. Tại bước trích xuất đặc trưng, server xử lý sử dụng bộ phát hiện đặc trưng Hessian-Affine<sup>[58]</sup> detector và bộ mô tả đặc trưng SIFT descriptor<sup>[30]</sup>, cả hai đều phụ thuộc phần lớn vào texture trên hình ảnh là các góc cạnh, đường viền. Việc giảm dung lượng hình ảnh như đã trình bày ở trên không làm mất các góc cạnh đường viền trên hình, vì vậy không ảnh hưởng nhiều tới kết quả truy vấn, mặt khác nó còn giúp tăng tốc độ xử lý trích xuất đặc trưng tại phía server.

- **Tại web service: giảm số lượng request bằng kỹ thuật long-polling**  
Kỹ thuật long-polling như đã trình bày trong mục [5.2.3](#) giúp giảm lượng request từ mỗi client khi có yêu cầu truy vấn hình ảnh. Vì vậy hệ thống có thể đáp ứng được nhiều request hơn, tốc độ phản hồi tới client vì thế cũng sẽ nhanh hơn.
- **Tại server xử lý: sử dụng phương pháp đề xuất và tối ưu kết quả trả về**

## **5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh**

---

Để tăng tốc độ truy vấn, hệ thống sử dụng phương pháp đề xuất đã trình bày trong mục [3.2](#) cùng những kết quả thí nghiệm chứng minh tính hiệu quả của phương pháp trong chương [4](#). Cùng với đó là việc tối ưu kết quả trả về để giảm tối đa thời gian phản hồi tới client.

Kết quả của quá trình truy vấn là một danh sách các hình ảnh và việc trả về những hình ảnh này cho client cũng sẽ chịu ảnh hưởng rất lớn từ tốc độ băng thông. Với một yêu cầu truy vấn, server chỉ cần trả về một số hình ảnh đầu trong danh sách xếp hạng là đầu, và chỉ trả về các hình ảnh thu nhỏ để giảm tối đa thời gian phản hồi về client. Cụ thể hơn, khi có kết quả truy vấn, server sẽ trả về cho client 15 hình ảnh thu nhỏ đúng đầu danh sách xếp hạng, mỗi hình thu nhỏ này có kích thước tối đa là 150 x 150 pixel và kích cỡ trung bình khoảng 5kB mỗi hình. Như vậy, tổng dung lượng cần gửi về client cho mỗi lần truy vấn là khoảng 75kB, vẫn có thể đáp ứng trên những kết nối chậm.

Thêm vào đó, việc sử dụng các server truy xuất dữ liệu độc lập với quá trình xử lý truy vấn cũng giúp tăng hiệu suất của hệ thống bằng cách phân hồi từng phần dữ liệu về client. Với chức năng xem cụ thể hình ảnh trả về, các server truy xuất dữ liệu sẽ nhận yêu cầu xem hình ảnh từ client và gửi về một hình ảnh thu nhỏ với kích thước tối đa là 500 x 500 pixel, với dung lượng khoảng 30kB mỗi hình, vẫn có thể hiển thị tốt trên điện thoại di động màn hình Full HD (1920x1080) trong khi dung lượng giảm tới 95% (với bộ dữ liệu oxford 5k, mỗi hình ảnh gốc nặng trung bình 500kB).

Hiệu quả của việc tối ưu các thành phần trên sẽ được kiểm chứng trong phần tiếp theo.

### **5.3 Cài đặt và thực nghiệm**

#### **5.3.1 Môi trường cài đặt**

Môi trường cài đặt của hệ thống được liệt kê chi tiết dưới đây:

- Processing Server:

## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh

---

- Hệ điều hành: Windows Server 2008.
- Ngôn ngữ lập trình: Matlab, C, Java.

- **Web Service:**

- Web server: Apache Tomcat 7.
- Ngôn ngữ lập trình: Java.

- **Client:**

- Hệ điều hành: Android 4.0.
- Ngôn ngữ lập trình: Java.

### 5.3.2 Kết quả thực nghiệm và đánh giá ứng dụng

Chúng tôi đánh giá ứng dụng dựa trên hai yếu tố chính: giao diện người dùng và hiệu suất truy vấn.

Về giao diện người dùng, ứng dụng thể hiện tính thân thiện cao và dễ sử dụng, cung cấp cho người dùng đủ các tính năng cơ bản của một hệ thống tìm kiếm hình ảnh.

Để đánh giá hiệu suất truy vấn, chúng tôi đo **thời gian phản hồi** của một truy vấn cơ bản, tính từ khi client gửi yêu cầu truy vấn tới khi nhận được kết quả, thực hiện trên bộ dữ liệu chuẩn Oxford Buildings 5k (Bảng 4.1.1.3).Thêm vào đó, để đánh giá chi tiết hơn, chúng tôi cũng đo các thông số sau:

- Dung lượng (DL) hình ảnh truy vấn sau mã hóa (MH).
- Thời gian xử lý tại client và server: tại client, thời gian xử lý là thời gian nén và mã hóa hình ảnh truy vấn và giải mã kết quả. Tại server, thời gian xử lý bao gồm: trích xuất đặc trưng (TXDT), truy vấn, giải mã (GM) và mã hóa (MH) và truy xuất dữ liệu (TXDL).
- Tổng thời gian truyền tải: Tổng chi phí thời gian truyền tải dữ liệu giữa client, web service và server xử lý cho một truy vấn cơ bản.
- Tổng dung lượng kết quả trả về.

Bảng 5.12 thể hiện kết quả thực nghiệm trên 10 truy vấn với các hình ảnh khác nhau. Có thể thấy tổng thời gian phản hồi của một truy vấn phụ thuộc rất lớn vào thời gian xử lý tại server, cụ thể hơn là thời gian trích xuất đặc trưng.

## 5. Xây dựng ứng dụng tìm kiếm đối tượng trên ảnh

---

STT	DL Hình ảnh truy vấn Sau MH	Thời gian xử lý (giây)					Tổng thời gian truyền tải (giây)	Tổng thời gian phản hồi (giây)	Dung lượng kết quả			
		Tại Client	Tại Server									
			TXDT	Truy vấn	GM +MH +TXDL	Tổng						
1	21.51 kB	0.065	3.217	0.008	0.031	3.256	1.650	4.961	50.58 kB			
2	16.93 kB	0.042	2.934	0.009	0.028	2.342	1.241	3.583	65.55 kB			
3	13.45 kB	0.045	1.725	0.01	0.053	1.788	1.160	2.948	54.13 kB			
4	40.25 kB	0.052	11.91	0.018	0.074	12.00	1.240	13.240	77.24 kB			
5	21.00 kB	0.035	2.198	0.008	0.031	2.237	0.509	2.746	56.06 kB			
6	34.32 kB	0.090	6.792	0.01	0.037	6.838	0.962	7.800	53.29 kB			
7	40.46 kB	0.102	7.780	0.015	0.037	7.832	1.820	9.652	64.81 kB			
8	21.75 kB	0.042	2.414	0.009	0.055	2.478	1.532	4.001	59.89 kB			
9	14.30 kB	0.089	1.953	0.011	0.068	2.032	0.545	2.577	46.31 kB			
10	25.84 kB	0.073	4.499	0.011	0.048	4.558	0.988	5.546	78.31 kB			
<b>AVG</b>	24.98 kB	0.0635	<b>4.542</b>	0.010	0.0462	4.536	1.164	<b>5.706</b>	60.62kB			

Hình 5.12: Kết quả thực nghiệm hiệu suất của hệ thống tìm kiếm đối tượng trên ảnh.

Với kết quả trên, thời gian trích xuất đặc trưng trung bình là 4.542 giây, chiếm 79.6% tổng thời gian phản hồi. Trong khi đó, thời gian truyền tải cho toàn bộ quá trình là 1.164 giây, với dữ liệu gồm nhiều hình ảnh thì kết quả này là khá tốt, thể hiện hiệu suất của ứng dụng tại các bước nén và mã hóa hình ảnh, tối ưu quá trình truyền dẫn. Với tổng lưu lượng truyền dẫn cho cả quá trình là 85.6kB (dung lượng hình ảnh truy vấn và tổng dung lượng kết quả trả về) hoàn toàn có thể đáp ứng những kết nối chậm, chẳng hạn mạng EDGE với tốc độ 236kb/s.

Ngoài ra, một yếu tố quan trọng để đánh giá hiệu suất của ứng dụng là độ chính xác truy vấn. Ứng dụng sử dụng phương pháp đề xuất đã được trình bày trong mục 3 và kết quả đã được thí nghiệm trên các bộ dữ liệu chuẩn tại mục 4.

Như vậy, những thực nghiệm trên cho thấy ứng dụng có thể đáp ứng được các yêu cầu thực tế của một hệ thống tìm kiếm đối tượng trên ảnh.

# Chương 6

## Kết luận và hướng phát triển

### 6.1 Kết luận

Với những kiến thức cơ sở và sự tìm hiểu, nghiên cứu các công trình trong lĩnh vực truy vấn ảnh, chúng tôi đã hệ thống lại những nền tảng kiến thức quan trọng. Từ đó, đề xuất phương pháp cải tiến nhằm nâng cao hiệu suất của hệ thống truy vấn ảnh trên tập dữ liệu lớn phục vụ cho các ứng dụng thời gian thực.

Để đánh giá hiệu quả của phương pháp đề xuất, chúng tôi đã tiến hành cài đặt và thử nghiệm với ba bộ dữ liệu chuẩn là Oxford 5K, Paris 6K và Oxford 100K đồng thời so sánh với các phương pháp cơ bản phổ biến hiện nay. Kết quả thí nghiệm được đánh giá theo quy trình đánh giá chuẩn được dùng cho các hệ thống truy vấn ảnh. Kết quả đạt được cho thấy phương pháp đề xuất đã giúp nâng cao hiệu suất của hệ thống truy vấn và đạt được sự cân bằng giữa độ chính xác và thời gian truy vấn.

Kết quả nghiên cứu này đã được tổng hợp thành bài báo gửi đăng tại hội nghị *The IEEE International Symposium on Multimedia* (ISM2014): Bien-Van Nguyen, Duy Pham, Thanh Duc Ngo, Duy-Dinh Le and Anh Duc Duong, “**Integrating Spatial Information into Inverted Index for Large-Scale Image Retrieval**”.

Cùng với đó, từ những kiến thức nền tảng và phương pháp cải tiến, chúng tôi đã xây dựng được một ứng dụng thực nghiệm đáp ứng được các yêu cầu về độ chính xác và thời gian phản hồi, là tiền đề để xây dựng những ứng dụng giải quyết những vấn đề cụ thể trong thực tế.

---

## 6.2 Hướng phát triển

### 6.2.1 Mở rộng phương pháp đề xuất

Để có thể xây dựng được những hệ thống truy vấn ảnh ứng dụng trong thực tế có khả năng truy vấn trên cơ sở dữ liệu gồm hàng triệu hoặc thậm chí hàng tỉ hình ảnh trong thời gian thực, sẽ cần rất nhiều thứ cần làm và ta cũng không thể nào biết được như thế nào sẽ là đủ để cho ra đời một hệ thống đáp ứng được các yêu cầu trong thực tế. Dưới đây chúng tôi chỉ nêu ra một vài hướng mở rộng cho công trình này.

**Cải tiến phương pháp xếp hạng.** Phương pháp xếp hạng bầu chọn (voting) chúng tôi dùng trong công trình này vẫn còn khá sơ khai và chưa tận dụng hết được thông tin không gian ảnh của chỉ mục ngược. Cụ thể, phương pháp bầu chọn mới chỉ quan tâm tới việc hai ô vuông trong không gian phân cấp có chứa cùng một visual word hay không chứ không quan tâm tới con số của visual word đó chứa trong mỗi ô. Đồng thời cũng phải quan tâm tới việc đánh trọng số cho trường hợp này để tránh rơi vào trường hợp có quá nhiều visual word giống nhau tập trung trong một ô cục bộ.

**Kết hợp với các phương pháp xếp hạng khác.** Phương pháp của chúng tôi hoàn toàn có thể kết hợp với những phương pháp xếp hạng khác như **tf-idf** để tăng kết quả truy vấn bằng cách xếp hạng lại một phần hoặc thậm chí dùng làm trọng số để xếp hạng lại toàn bộ kết quả. Chẳng hạn phương pháp Spatial Pyramid Matching thường dùng để xếp hạng lại kết quả của **tf-idf**.

**Thay đổi cấu trúc của chỉ mục ngược.** Cấu trúc của chỉ mục ngược vẫn chỉ dừng lại ở việc lưu trữ danh sách hình ảnh có chứa một visual word nào đó, do đó vẫn chưa tận dụng hết được khả năng của chỉ mục ngược. Ta có thể mở rộng cấu trúc của chỉ mục ngược để phục vụ cho việc lưu trữ các thông tin khác như trọng số tương ứng của từng visual word, số lượng của visual word đó trong ảnh,...

### 6.2.2 Phát triển ứng dụng thực tế

Ứng dụng tìm kiếm đối tượng trên ảnh chúng tôi xây dựng đã chứng minh được tính thực tiễn của công trình nghiên cứu này, đồng thời mở ra hướng ứng dụng

---

trên nhiều lĩnh vực khác nhau trong thực tế.

Phát triển một ứng dụng giải quyết một vấn đề cụ thể trong đời sống. Chúng tôi hướng tới một ứng dụng cụ thể hơn, có thể là ứng dụng tìm kiếm thông tin sản phẩm, quảng bá sản phẩm theo ngữ cảnh, bảo vệ thương hiệu hay một ứng dụng tìm kiếm địa điểm du lịch... Sử dụng kết quả của hệ thống cơ bản làm nền tảng để phân tích, kết hợp với nhiều thông tin khác nhau trong từng lĩnh vực, từ đó đưa ra được kết quả cần thiết cho từng bài toán cụ thể.

Hơn nữa, với việc thiết kế một ứng dụng thân thiện với người dùng và việc tối ưu hệ thống để đáp ứng được nhiều yêu cầu cùng lúc, cùng với khả năng chia tải để xử lý, chúng tôi hướng tới một ứng dụng thực tế đáp ứng được lượng người dùng lớn với thời gian phản hồi nhanh.

# Tài liệu tham khảo

- [1] I. T. Young, J. E. Walker, and J. E. Bowie, “An analysis technique for biological shape. i,” *Information and control*, vol. 25, no. 4, pp. 357–370, 1974. [11](#)
- [2] M. Peura and J. Iivarinen, “Efficiency of simple shape descriptors,” in *Proceedings of the third international workshop on visual form*, vol. 443. Citeseer, 1997, p. 451. [11](#)
- [3] D. Chetverikov and Y. Khenokh, “Matching for shape defect detection,” in *Computer Analysis of Images and Patterns*. Springer, 1999, pp. 367–374. [11](#)
- [4] E. R. Davies, *Machine vision: theory, algorithms, practicalities*. Elsevier, 2004. [11](#)
- [5] P. J. Van Otterloo, *A contour-oriented approach to shape analysis*. Prentice Hall International (UK) Ltd., 1991. [11](#)
- [6] D. Zhang, G. Lu *et al.*, “A comparative study of fourier descriptors for shape representation and retrieval,” in *Proc. of 5th Asian Conference on Computer Vision (ACCV)*. Citeseer, 2002, pp. 652–657. [11](#)
- [7] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*. Cengage Learning, 2014. [11](#)
- [8] R. C. Gonzalez and R. E. Woods, “Digital image processing,” pp. 502–503, 2002. [11](#)

## TÀI LIỆU THAM KHẢO

---

- [9] A. Del Bimbo and P. Pala, “Visual image retrieval by elastic matching of user sketches,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 19, no. 2, pp. 121–132, 1997. [11](#)
- [10] H. Asada and M. Brady, “The curvature primal sketch,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 1, pp. 2–14, 1986. [11](#)
- [11] T. Pavlidis, *Algorithms for graphics and image processing*. Computer science press, 1982. [11](#)
- [12] M.-K. Hu, “Visual pattern recognition by moment invariants,” *Information Theory, IRE Transactions on*, vol. 8, no. 2, pp. 179–187, 1962. [12](#)
- [13] G. Taubin and D. B. Cooper, “Recognition and positioning of rigid objects using algebraic moment invariants,” in *San Diego,’91, San Diego, CA*. International Society for Optics and Photonics, 1991, pp. 175–186. [12](#)
- [14] ——, *Object recognition based on moment (or algebraic) invariants*. IBM TJ Watson Research Center, 1991. [12](#)
- [15] M. R. Teague, “Image analysis via the general theory of moments\*,” *JOSA*, vol. 70, no. 8, pp. 920–930, 1980. [12](#)
- [16] T. Meier and K. N. Ngan, “Automatic segmentation of moving objects for video object plane generation,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 8, no. 5, pp. 525–538, 1998. [12](#)
- [17] M. Tuceryan and A. K. Jain, “Texture analysis,” *The handbook of pattern recognition and computer vision*, vol. 2, pp. 207–248, 1998. [12](#), [13](#)
- [18] B. S. Manjunath and W.-Y. Ma, “Texture features for browsing and retrieval of image data,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 18, no. 8, pp. 837–842, 1996. [12](#), [13](#)
- [19] J. A. Montoya Zegarra, N. J. Leite, and R. da Silva Torres, “Wavelet-based fingerprint image retrieval,” *Journal of computational and applied mathematics*, vol. 227, no. 2, pp. 294–307, 2009. [12](#)

## TÀI LIỆU THAM KHẢO

---

- [20] F. Xu and Y.-J. Zhang, “Evaluation and comparison of texture descriptors proposed in mpeg-7,” *Journal of Visual Communication and Image Representation*, vol. 17, no. 4, pp. 701–716, 2006. [12](#)
- [21] H. Tamura, S. Mori, and T. Yamawaki, “Textural features corresponding to visual perception,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 8, no. 6, pp. 460–473, 1978. [12](#)
- [22] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, “Textural features for image classification,” *Systems, Man and Cybernetics, IEEE Transactions on*, no. 6, pp. 610–621, 1973. [12](#)
- [23] R. Nevatia and R. Nevatia, *Machine perception*. Prentice-Hall Englewood Cliffs, NJ, 1982. [13](#)
- [24] P. Wu, B. Manjunath, S. Newsam, and H. Shin, “A texture descriptor for browsing and similarity retrieval,” *Signal Processing: Image Communication*, vol. 16, no. 1, pp. 33–43, 2000. [13](#)
- [25] M. J. Swain and D. H. Ballard, “Color indexing,” *International journal of computer vision*, vol. 7, no. 1, pp. 11–32, 1991. [13](#), [14](#)
- [26] M. A. Stricker and M. Orengo, “Similarity of color images,” in *IS&T/SPIE’s Symposium on Electronic Imaging: Science & Technology*. International Society for Optics and Photonics, 1995, pp. 381–392. [13](#)
- [27] R. O. Stehling, M. A. Nascimento, and A. X. Falcão, “Cell histograms versus color histograms for image representation and retrieval,” *Knowledge and Information Systems*, vol. 5, no. 3, pp. 315–336, 2003. [14](#)
- [28] Y. Deng, B. Manjunath, C. Kenney, M. S. Moore, and H. Shin, “An efficient color representation for image retrieval,” *Image Processing, IEEE Transactions on*, vol. 10, no. 1, pp. 140–147, 2001. [14](#)
- [29] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, “Color and texture descriptors,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 11, no. 6, pp. 703–715, 2001. [14](#)

## TÀI LIỆU THAM KHẢO

---

- [30] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004. [15](#), [62](#)
- [31] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004. [15](#)
- [32] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” *International journal of computer vision*, vol. 60, no. 1, pp. 63–86, 2004. [15](#)
- [33] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 404–417. [15](#)
- [34] E. Rosten, R. Porter, and T. Drummond, “Faster and better: A machine learning approach to corner detection,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 1, pp. 105–119, 2010. [15](#)
- [35] S. Leutenegger, M. Chli, and R. Y. Siegwart, “Brisk: Binary robust invariant scalable keypoints,” in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2548–2555. [15](#)
- [36] K. Mikolajczyk and C. Schmid, “A performance evaluation of local descriptors,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 10, pp. 1615–1630, 2005. [15](#)
- [37] E. Tola, V. Lepetit, and P. Fua, “A fast local descriptor for dense matching,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8. [15](#)
- [38] Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven, “Tour the world: building a web-scale landmark recognition engine,” in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1085–1092. [15](#)

## TÀI LIỆU THAM KHẢO

---

- [39] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, “Brief: Binary robust independent elementary features,” in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 778–792. [15](#)
- [40] R. Arandjelovic and A. Zisserman, “Three things everyone should know to improve object retrieval,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2911–2918. [15, 16, 35](#)
- [41] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 1470–1477. [16, 17, 22](#)
- [42] ——, “Efficient visual search of videos cast as text retrieval,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 4, pp. 591–606, 2009. [18](#)
- [43] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Object retrieval with large vocabularies and fast spatial matching,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. IEEE, 2007, pp. 1–8. [17, 18, 22, 32, 33, 35, 36, 37, 49](#)
- [44] D. Nister and H. Stewenius, “Scalable recognition with a vocabulary tree,” in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 2161–2168. [17](#)
- [45] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1. [18](#)
- [46] E. Krauss, “Taxicab geometry: an adventure in non-euclidean geometry,” 1987. [19](#)
- [47] M. M. Deza and E. Deza, *Encyclopedia of distances*. Springer, 2009. [19](#)
- [48] E. Cheng, N. Xie, H. Ling, P. R. Bakic, A. D. Maidment, and V. Megalooikonomou, “Mammographic image classification using histogram intersection,” in *Biomedical Imaging: From Nano to Macro, 2010 IEEE International Symposium on*. IEEE, 2010, pp. 197–200. [19, 20](#)

## TÀI LIỆU THAM KHẢO

---

- [49] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong, “A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry,” *Artificial intelligence*, vol. 78, no. 1, pp. 87–119, 1995. [22](#)
- [50] C. Schmid, R. Mohr *et al.*, “Local grayvalue invariants for image retrieval,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 5, pp. 530–534, 1997. [22](#)
- [51] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981. [22](#)
- [52] Y. Zhang, Z. Jia, and T. Chen, “Image retrieval with geometry-preserving visual phrases,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 809–816. [22](#)
- [53] Z. Lin and J. Brandt, “A local bag-of-features model for large-scale object retrieval,” in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 294–308. [22](#)
- [54] C. H. Lampert, “Detecting objects in large image collections and videos by efficient subimage retrieval,” in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 987–994. [22](#)
- [55] S. Lazebnik, C. Schmid, and J. Ponce, “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. IEEE, 2006, pp. 2169–2178. [23](#), [25](#), [27](#), [28](#), [35](#)
- [56] K. Grauman and T. Darrell, “The pyramid match kernel: Discriminative classification with sets of image features,” in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1458–1465. [23](#), [28](#)

## TÀI LIỆU THAM KHẢO

- [57] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, “Lost in quantization: Improving particular object retrieval in large scale image databases,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on.* IEEE, 2008, pp. 1–8. [32](#)
- [58] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schafalitzky, T. Kadir, and L. Van Gool, “A comparison of affine region detectors,” *International journal of computer vision*, vol. 65, no. 1-2, pp. 43–72, 2005. [35, 62](#)