

**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 CÁM ƠN

Tôi xin chân thành cảm ơn ...

## TÓM TẮT

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

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

Trong luận văn này, chúng tôi đề xuất một phương pháp nhằm tích hợp thông tin không gian ảnh vào phương pháp đánh chỉ mục ngược

(Inverted Index) để nâng cao độ chính xác nhưng vẫn đảm bảo được thời gian truy vấn nhanh. Kết quả thí nghiệm trên các tập dữ liệu chuẩn như Oxford 5k, Paris 6k và Holiday đã cho thấy tính hiệu quả của phương pháp này.

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

# Mục lục

<b>Mục lục</b>	<b>iv</b>
<b>Danh sách hình vẽ</b>	<b>vii</b>
<b>Danh sách bảng</b>	<b>viii</b>
<b>Danh sách từ viết tắt</b>	<b>ix</b>
<b>1 Tổng quan</b>	<b>1</b>
1.1 Đặt vấn đề . . . . .	1
1.1.1 Một vài hướng ứng dụng của hệ thống truy vấn ảnh . . . . .	2
1.2 Thách thức . . . . .	4
1.3 Mục đích, đối tượng và phạm vi nghiên cứu . . . . .	5
1.3.1 Mục đích . . . . .	5
1.3.2 Đối tượng nghiên cứu . . . . .	6
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 Biểu diễn hình ảnh bằng các đặc trưng cục bộ . . . . .	9
2.2 Mô hình Bag-of-words . . . . .	10
2.2.1 Truy vấn văn bản . . . . .	11
2.2.2 Bag-of-Words trong truy vấn ảnh . . . . .	12
2.3 Sử dụng thông tin không gian ảnh trong truy vấn ảnh . . . . .	14
2.3.1 Các hướng tiếp cận dựa trên đặc trưng hình học . . . . .	15

## MỤC LỤC

---

2.3.2 Các hướng tiếp cận dựa trên thông tin không gian của các đặc trưng cục bộ . . . . .	16
2.4 Kết chương . . . . .	16
<b>3 Phương pháp đề xuất</b>	<b>17</b>
3.1 Chỉ mục ngược với biểu diễn Bag-of-Visual-Words . . . . .	17
3.2 Tích hợp thông tin không gian ảnh vào chỉ mục ngược . . . . .	19
3.3 Cải thiện hiệu suất của hệ thống . . . . .	21
3.3.1 Tăng độ chính xác của quá trình gom cụm . . . . .	23
3.3.2 Lọc bỏ các stop word . . . . .	24
<b>4 Thực nghiệm và đánh giá kết quả</b>	<b>25</b>
4.1 Các bộ dữ liệu và phương thức đánh giá . . . . .	25
4.1.1 Các bộ dữ liệu . . . . .	26
4.1.1.1 Oxford 5K . . . . .	26
4.1.1.2 Paris 6k . . . . .	26
4.1.1.3 Oxford 5K+100K . . . . .	26
4.1.2 Phương thức đánh giá . . . . .	28
4.2 Cài đặt thí nghiệm . . . . .	29
4.3 Kết quả thí nghiệm và đánh giá kết quả . . . . .	30
<b>5 Xây dựng ứng dụng</b>	
<b>tìm kiếm đối tượng trên ảnh</b>	<b>37</b>
5.1 Tổng quan ứng dụng . . . . .	37
5.1.1 Mục đích và phạm vi của ứng dụng . . . . .	37
5.1.2 Các chức năng chính . . . . .	38
5.2 Thiết kế ứng dụng . . . . .	38
5.2.1 Kiến trúc . . . . .	38
5.2.1.1 Client side . . . . .	39
5.2.1.2 Web service . . . . .	39
5.2.1.3 Server side . . . . .	39
5.2.2 Giao diện . . . . .	40
5.3 Cài đặt và thử nghiệm . . . . .	40
5.3.1 Môi trường cài đặt . . . . .	40

## **MỤC LỤC**

---

5.3.2	Kết quả thử nghiệm . . . . .	40
5.3.3	Dánh giá kết quả . . . . .	40
<b>6</b>	<b>Tổng kết</b>	<b>41</b>
6.1	Kết luận . . . . .	41
6.2	Hướng phát triển . . . . .	41
	<b>Tài liệu tham khảo</b>	<b>43</b>

# Danh sách hình vẽ

1.1	Sơ đồ tổng quát của một hệ thống truy vấn ảnh . . . . .	2
1.2	Những thay đổi bề ngoài của đối tượng trên ảnh . . . . .	4
2.1	Các từ trực quan (visual words) . . . . .	13
2.2	Bỏ qua thông tin không gian ảnh trong mô hình BoW . . . . .	14
3.1	Quá trình tạo tập chỉ mục ngược . . . . .	18
3.2	Khái quát về phương pháp đề xuất . . . . .	21
3.3	Quá trình truy vấn của phương pháp đề xuất . . . . .	22
4.1	Landmark và các truy vấn được dùng để đánh giá . . . . .	27
4.2	Biểu đồ so sánh độ chính xác và thời gian truy vấn giữa các phương pháp . . . . .	31
4.3	Biểu đồ so sánh độ chính xác và thời gian truy vấn giữa các phương pháp . . . . .	33
4.4	Ví dụ thể hiện kết quả khi tìm kiếm một hình ảnh trên bộ dữ liệu Oxford 5k . . . . .	34
4.5	Ví dụ thể hiện kết quả khi tìm kiếm một hình ảnh trên bộ dữ liệu Paris 6k . . . . .	36
5.1	Kiến trúc tổng quát của hệ thống . . . . .	39

# Danh sách bảng

3.1	So sánh kết quả truy vấn với số lần lặp khác nhau trong thuật toán gom cụm AKM . . . . .	23
3.2	Kết quả lọc bỏ các stop words . . . . .	24
4.1	Số hình ảnh trong mỗi bộ dữ liệu và số truy vấn trong tập dữ liệu đánh giá chuẩn tương ứng . . . . .	28
4.2	Hiệu suất của các phương pháp trên bộ dữ liệu Oxford 5k . . . . .	30
4.3	Hiệu suất của các phương pháp trên bộ dữ liệu Paris 6k . . . . .	31
4.4	Hiệu suất của các phương pháp trên bộ dữ liệu Holidays . . . . .	32
4.5	Hiệu suất của phương pháp đề xuất với các giá trị khác nhau của $L$ trên bộ dữ liệu Oxford 5K . . . . .	33
4.6	Hiệu suất của phương pháp đề xuất với các giá trị khác nhau của $L$ trên bộ dữ liệu Paris 6K . . . . .	35

# 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

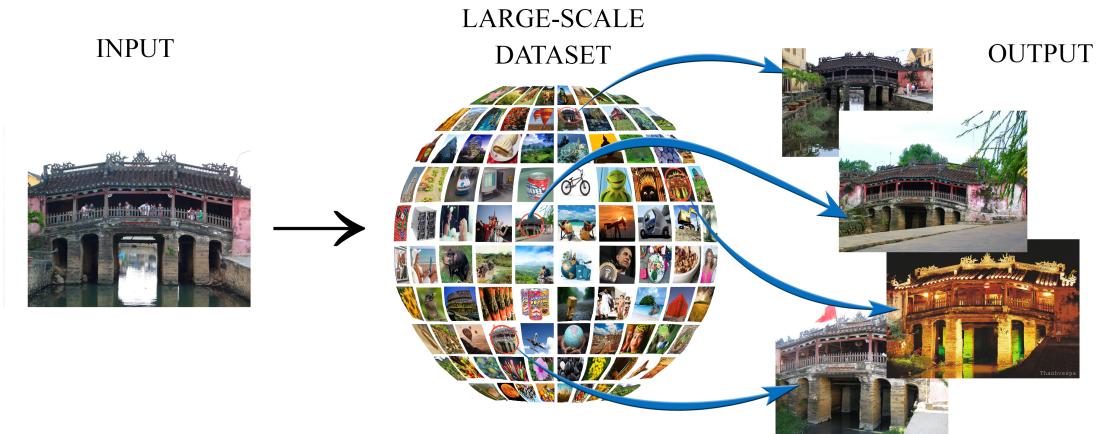
## Tổng quan

### 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 thông tin số khổng lồ bao gồm hình ảnh, video, v.v... Do đó, nhu cầu truy vấn thông tin từ kho dữ liệu hình ảnh, video ngày càng bức thiết hơn bao giờ hết.

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

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



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

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

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

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

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

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

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

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

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

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

## 1.2 Thách thức

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

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



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

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

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

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

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

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

#### 1.3.1 Mục đích

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 và kết quả phải được trả về gần như ngay lập tức với cơ sở dữ liệu gồm hàng triệu hình ảnh chưa được gán nhãn. Hệ thống này tập trung vào giải quyết vấn đề về tìm kiếm một đối tượng cụ thể như một địa điểm, một bức tranh, một bìa sách, v.v... Những đối tượng này có thể được chụp trong các điều kiện khác nhau như góc chụp, ánh sáng, kích thước hay bị che khuất. Do đó mục đích của hệ thống không phải là trả về những hình ảnh chụp gần giống nhau như

chụp trong cùng một khung cảnh hay cùng thuộc một loại đối tượng mà là trả về những hình ảnh có chứa chính xác đối tượng cần tìm. Ví dụ như khi đưa vào một bức hình có chứa Nhà thờ Đức Bà, kết quả trả về sẽ những bức hình có chứa nhà thờ Đức Bà chứ không phải trả về những nhà thờ có kiến trúc hay có 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 vấn đề như: 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**

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

- Các kiến thức nền tảng trong lĩnh vực xử lý ảnh như dò tìm và phát hiện các đặc trưng ảnh, rút trích đặc trưng,...
- Ứng dụng một vài kỹ thuật trong lĩnh vực phân lớp văn bản vào trong xử lý ảnh như mô hình Bag-of-Words (BoW), phương pháp tf-idf,...
- Các công trình nghiên cứu liên quan đến bài toán tìm kiếm và phân lớp 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 quát về các công trình nghiên cứu liên quan tới truy vấn ảnh và bàn luận chi tiết về từng công trình trong Chương 2.

**Các tập dữ liệu và phương pháp đánh giá.** Để thử nghiệm kết quả của

## Chương 1. Tổng quan

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

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

**Tổng kết.** Trong Chương 6, chúng tôi sẽ tổng kết, bàn luận thêm về phương pháp đề xuất và những đề xuất cải tiến, mở rộng để nâng cao hiệu suất của hệ thống trong thời gian tới.

## Chương 2

# Các công trình liên quan

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

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

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

## 2.1 Biểu diễn hình ảnh bằng các đặc trưng cục bộ

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

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

Có rất nhiều phương pháp dò tìm các đặc trưng (feature detector) được đưa ra, trong đó phải kể tới các phương pháp được dùng phổ biến như Difference of Gaussians, DoG [1], Maximally Stable Extremal Regions, MSER [2] và affine invariant detector [3]. Ngoài ra còn có các phương pháp dò tìm được xây dựng để tìm kiếm trong thời gian thực như SURF [4], FAST [5] và BRISK [6].

Sau khi rút trích được các đặc trưng cục bộ cho mỗi hình, dựa trên các đặc trưng đó ta sẽ quyết định xem liệu hai tấm hình bất kỳ có chứa cùng một đối tượng hay không. Để so sánh độ tương đồng của hai đặc trưng cục bộ, ta không thể dựa trên màu sắc và cường độ của chúng vì những yếu tố này không bền vững trước những thay đổi của hình ảnh. Do đó ta cần phải tìm cách lượng tử hóa độ tương đồng giữa cách đặc trưng để có thể đo được bằng các tính toán cụ thể. Trong công trình nghiên cứu nổi tiếng của Lowe [1], 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 [7], SURF [4], DAISY [8], CONGAS [9], BRIEF [10]. Đặc biệt, bằng việc đề xuất thuật toán RootSIFT được cải tiến từ SIFT, Arandjelovic và Zisserman [11] đã nâng hiệu suất của phương pháp SIFT lên đáng kể. Đây cũng là phương pháp được chúng tôi chọn dùng trong hệ thống của mình.

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

## 2.2 Mô hình Bag-of-words

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

### 2.2.1 Truy vấn văn bản

Tương tự như hình ảnh, để có thể thực hiện truy vấn với văn bản, văn bản được biểu diễn dưới dạng một mô hình không gian vector [12] hay còn được gọi là mô hình *túi từ* (Bag-of-Words), BoW [13]. Theo đó, mỗi văn bản được xem như là một tập hỗn độn (một túi) các từ và được biểu diễn dưới dạng một biểu đồ (histogram)  $N_w$ -chiều với  $N_w$  là số các từ của một ngôn ngữ. Vì giá trị của mỗi cột cột của biểu đồ bằng với số lần xuất hiện của từ tương ứng với cột đó trong văn bản nên phương pháp này còn được gọi là *trọng số tần suất từ* (term frequency weighting).

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

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

Cuối cùng, trọng số của mỗi từ trong mỗi văn bản được tính bằng cách lấy tích của tần suất từ (term frequency - tf) và tần số nghịch đảo văn bản (invert document frequency - idf). Trọng số đó được gọi là tf-idf [13] với công thức:

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

Dối với những từ xuất hiện với tần suất cực kỳ lớn (stop word), ta có thể lọc và loại bỏ toàn bộ để giảm bớt chi phí về không gian lưu trữ và thời gian thực thi.

Mức độ tương đồng giữa các văn bản sẽ được tính bằng công thức cosin áp

dụng cho trọng số tf-idf của chúng trong mô hình BoW. Thực tế mỗi văn bản chỉ chứa một lượng rất nhỏ so với số lượng các từ có trong ngôn ngữ, do vậy vector sinh ra khi biểu diễn bằng mô hình BoW sẽ rất thưa thớt. Để cho quá trình lưu trữ và truy vấn được hiệu quả, một cấu trúc dữ liệu sẽ được tính toán trước được gọi là *chỉ mục ngược* (inverted index) [13]. Chỉ mục ngược bao gồm một chuỗi các danh sách, mỗi danh sách tương ứng với một từ. Mỗi danh sách ghi lại những văn bản nào có chứa từ đó. Nhờ chỉ mục ngược, khi đưa vào một danh sách các từ rút từ văn bản truy vấn, ta có thể nhanh chóng lấy được danh sách các văn bản trong tập văn bản chứa các từ truy vấn đó. Từ đó có thể dễ dàng tính ra chỉ số tf-idf cho từng từ.

### 2.2.2 Bag-of-Words trong truy vấn ảnh

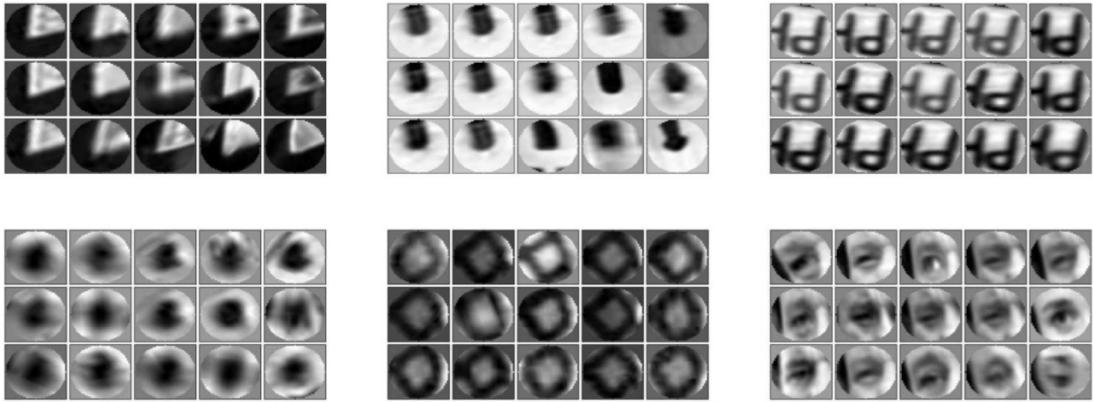
Một khó khăn lớn khi áp dụng mô hình của truy vấn văn bản vào truy vấn ảnh là trong truy vấn văn bản, một văn bản có thể dễ dàng bóc tách ra các từ trong khi đó không có cách phân chia tự nhiên nào cho các hình ảnh. Như đã giới thiệu trong mục 2.1, một hình ảnh hoàn toàn có thể chia thành các đặc trưng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

Nghiên cứu của Sivic và Zisserman [14] là công trình đầu tiên ứng dụng hướng tiếp cận của truy vấn văn bản vào truy vấn ảnh<sup>1</sup>. Trong công trình này tác giả đã giới thiệu khái niệm *các từ trực quan* (visual words) được tạo ra bằng cách sử dụng thuật toán gom cụm k-means để gom cụm các đặc trưng cục bộ. Hình 2.1 cho thấy một vài ví dụ về các từ trực quan. Tương tự như trong truy vấn văn bản, hình ảnh sẽ được rút trích các đặc trưng cục bộ rồi tiến hành gom cụm để biểu diễn thành các từ trực quan, sau đó được đánh trọng số bằng tf-idf, rồi biểu diễn dưới dạng mô hình BoW và sử dụng chỉ mục ngược để tăng hiệu suất cho quá trình truy vấn. Thí nghiệm được tiến hành trên 4000 ảnh (frame) được lấy từ video và rút trích được 10,000 từ trực quan từ những hình ảnh đó.

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

---

<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



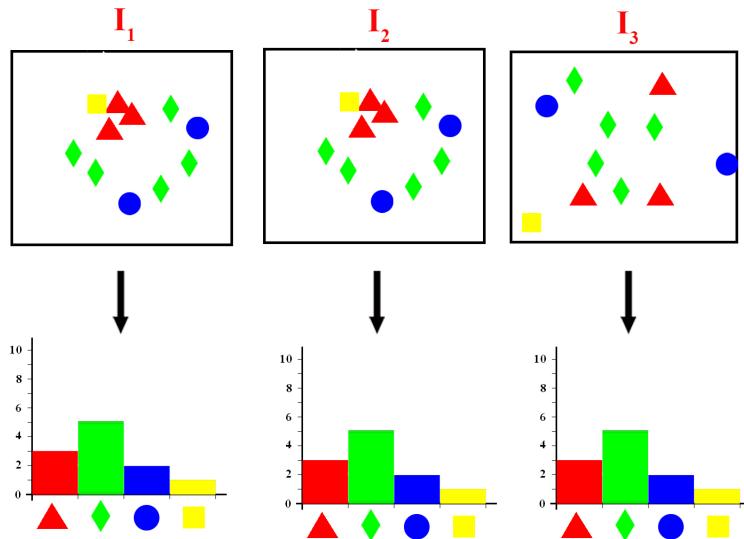
Hình 2.1: **Các từ trực quan (visual words).** Mỗi nhóm là một nhóm các đặc trưng cục bộ được rút trích từ hình ảnh, gom vào cùng một cụm và cùng được biểu diễn bằng một từ trực quan. Hình ảnh được lấy từ bài báo [15].

sẽ tăng lên, đồng thời nó cũng khiến cho tốc độ truy vấn nhanh hơn vì các biểu đồ BoW sẽ trở nên "thưa" hơn. Trong thực tế, để truy vấn ảnh trên những tập dữ liệu lớn, để cho kết quả tốt thì số lượng từ trực quan không thể vào khoảng 10,000 từ như trong thí nghiệm của Sivic và Zisserman [14] mà phải lên tới hàng triệu từ. Trong khi đó, độ phức tạp của thuật toán k-means là  $O(N_w N_d)$  với  $N_w$ ,  $N_d$  lần lượt là kích cỡ của từ trực quan 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 ta không thể dùng k-means cho bài toán này. Nister và Stewenius [16] đã đề xuất phương pháp giải quyết cho bài toán này bằng cách xây dựng một cây từ vựng mà về bản chất thì nó chính là thuật toán HKM (Hierarchical K-Means). Để minh họa cho thuật toán này, tác giả đã cho thử nghiệm trên bộ ảnh gồm 1 triệu hình ảnh. Không lâu sau đó, Philbin và các đồng nghiệp [17] đã đề xuất một hướng tiếp cận khác dựa trên thuật toán *xấp xỉ k-means*, AKM (Approximate K-Means). Tác giả cũng cho chạy thử nghiệm AKM trên 16.7 triệu đặc trưng để gom cụm thành 1 triệu từ. Các thí nghiệm cho thấy rằng, khi so sánh AKM với k-means thì về độ chính xác thì AKM xấp xỉ k-means tuy nhiên chi phí tính toán chỉ bằng một phần nhỏ của k-means. Còn khi so sánh AKM với HKM thì AKM không 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  $O(N_d \log(N_w))$ .

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

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



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

Để giải quyết vấn đề trên, rất nhiều công trình nghiên cứu đã được đưa ra.

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

### 2.3.1 Các hướng tiếp cận dựa trên đặc trưng hình học

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

Sivic và Zisserman [14] 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 [18] và [19].

Trong một công trình nghiên cứu [17], tác giả sử dụng thuật toán RANSAC [20] 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 [21] 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 [22] và [23] lại xếp hạng các hình ảnh dựa trên điểm số so khớp của hình ảnh truy vấn với những cửa sổ con được định vị trên hình. Phương pháp này mã hóa được nhiều thông tin không gian ảnh hơn so với một hình 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.

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

Hướng tiếp cận dựa trên đặc trưng hình học là hướng tiếp cận mang tính toàn cục, tức là xem xét đối tượng dưới một cái nhìn tổng quan, toàn thể chứ không xem xét chi tiết những thành phần cấu thành nó. Hướng tiếp cận dựa trên các đặc trưng cục bộ lại ngược lại, xem đối tượng là một tập hợp của nhiều thành phần và dựa trên những thành phần đó để xác định đối tượng. Lazebnik [24] đã giới thiệu một phương pháp nền tảng, được bắt nguồn từ ý tưởng *so khớp phân cấp* (pyramid matching) của Grauman và Darrell [25], 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.

## 2.4 Kết chương

Việc biểu diễn hình ảnh bằng các đặc trưng cục bộ đã đặt nền tảng cho việc đưa ra các phương pháp để truy vấn đối tượng trên ảnh. Mô hình 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 để nâng độ chính xác của truy vấn lên rất nhiều nhưng lại không quan tâm nhiều tới thời gian thực hiện.

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

# Chương 3

## Phương pháp đề xuất

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

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

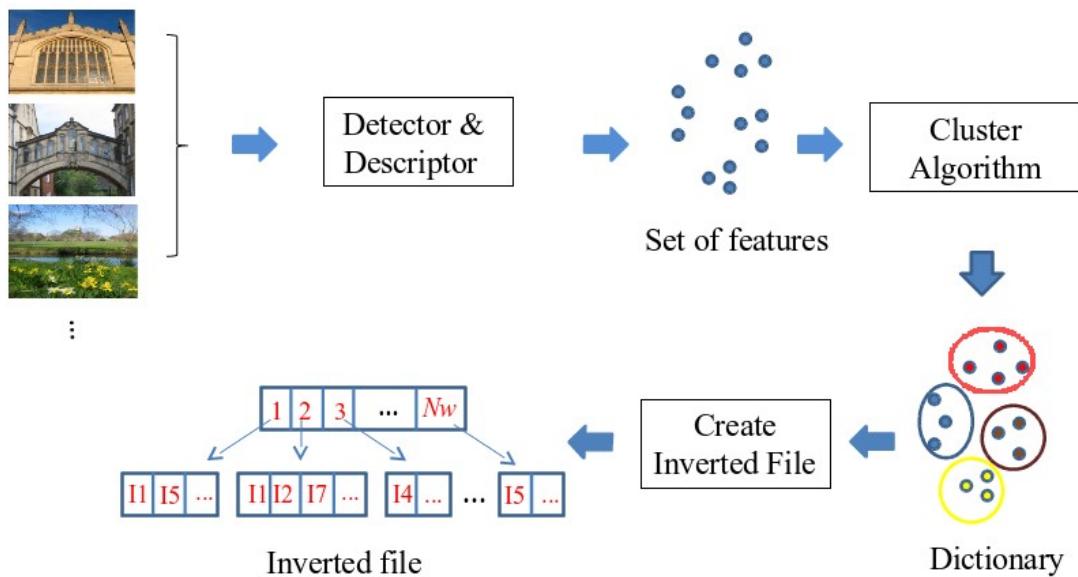
### 3.1 Chỉ mục ngược với biểu diễn Bag-of-Visual-Words

Như đã được giới thiệu sơ lược trong mục 2.2.1, chỉ mục ngược (inverted index) là phương pháp phổ dùng để tối ưu hóa tốc độ truy vấn cơ sở dữ liệu bằng việc

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

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. Mỗi khóa đại diện cho một *từ*, và phần giá trị của tương ứng lưu trữ danh sách các văn bản có chứa từ đó. Vì vậy ta có thể dễ dàng lấy được danh sách tất cả các văn bản chứa từ truy vấn.

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



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

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

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

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

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

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

Phương pháp chúng tôi đề xuất nhằm tích hợp thông tin không gian ảnh vào chỉ mục ngược được bắt nguồn từ ý tưởng của một công trình nghiên cứu của Lazebnik và các đồng nghiệp [24]. Trong công trình đó, thay vì sử dụng một biểu đồ (histogram) chung của các từ trực quan để biểu diễn một hình ảnh thì họ chia hình ảnh thành các nhiều phần sử dụng lưới ô vuông phân cấp (hay còn được gọi là không gian phân cấp - spatial pyramid). Một lưới ô vuông tại cấp  $l$  sẽ chia hình ảnh thành  $2^l \times 2^l$  ô với kích cỡ như nhau. Do đó, số ô vuông trên lưới ở cấp 0 là  $1 \times 1$ ; 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 hình BoW ta sẽ tính được các biểu đồ độc lập. Chính vì mức độ chia tiết của các biểu đồ khác nhau nên chúng sẽ được đánh trọng số khác rồi rồi được ghép nối với nhau để tạo thành một vector đặc trưng biểu diễn cho hình ảnh. Bằng cách biểu diễn như vậy, các hình ảnh có sự phân bố các từ

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

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

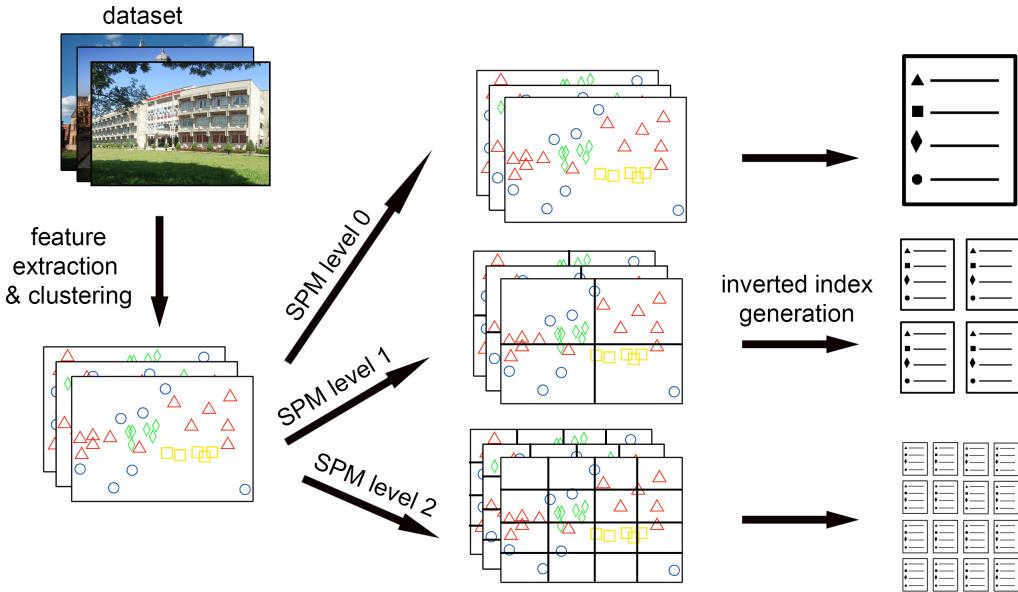
Dựa trên ý tưởng của nghiên cứu trên, chúng tôi đề xuất sử dụng không gian phân cấp để tăng cường mức độ bầu chọn và lập chỉ mục của kỹ thuật đánh chỉ mục ngược căn bản. Ý tưởng được chúng tôi đưa ra là chia hình ảnh thành nhiều ô sử dụng không gian phân cấp và giới hạn ở một cấp xác định. Sau đó các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

Một trong những điểm đặc biệt của phương pháp đề xuất là chúng tôi sử dụng đa chỉ mục ngược. Tức là chia thành nhiều tập chỉ mục ngược khác nhau nhưng các tập vẫn giữ được cấu trúc căn bản của chỉ mục ngược. Mỗi tập sẽ dùng để lưu trữ chỉ mục cho một ô trên không gian phân cấp. Nếu cấp độ cao nhất của không gian phân cấp là  $L$  thì tổng số lượng tập 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.

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

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

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



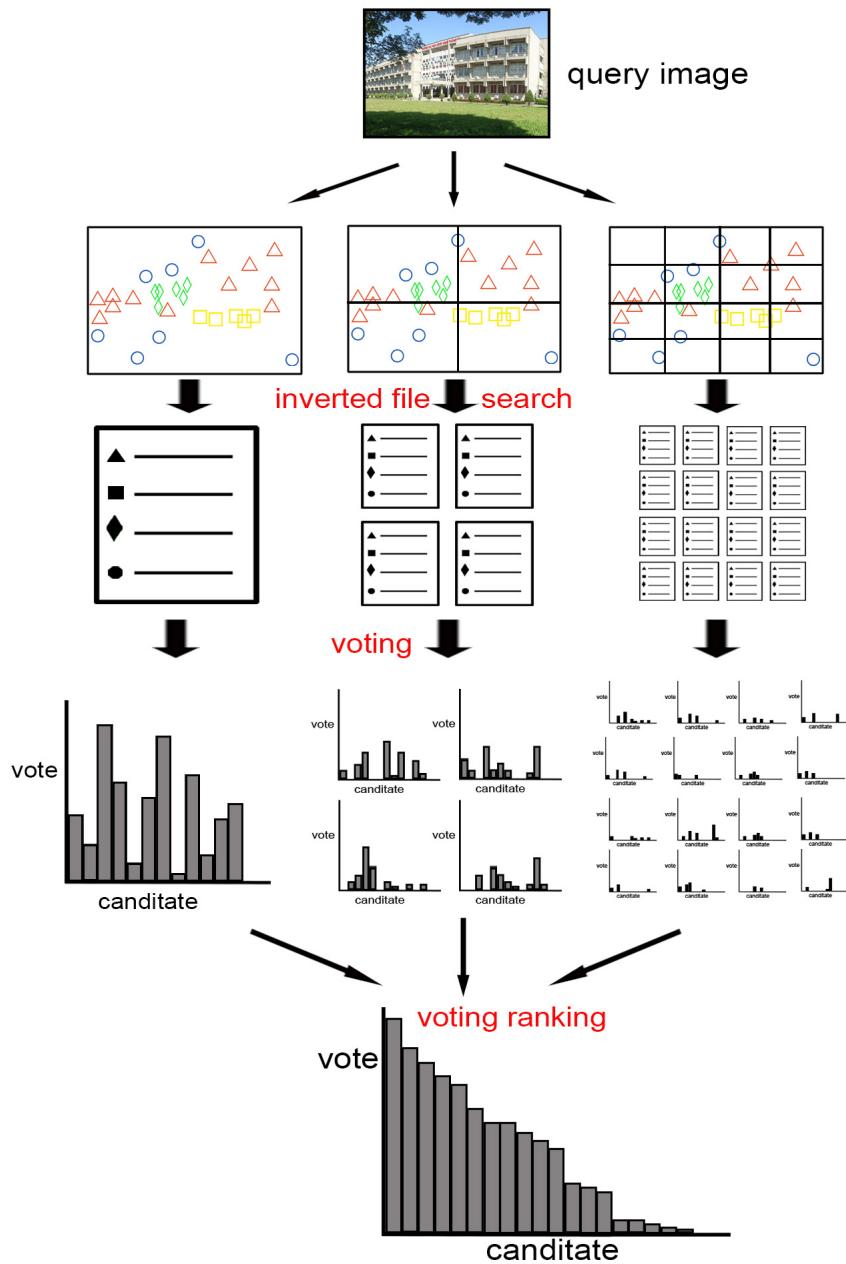
Hình 3.2: Khái quát về phương pháp đề xuất.

Dựa vào vị trí của các từ này, ta có thể xác định được chúng thuộc ô nào tại mỗi cấp của mô hình không gian phân cấp. Từ đó ta có thể có thể truy xuất ngay lập tức tới tập 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.

### 3.3 Cải thiện 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 do khoảng cách phù hợp

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



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

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

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

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.

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

Quá trình gom cụm các đặc trưng tạo thành các từ trực quan càng chính xác thì độ chính xác của quá trình truy vấn càng cao. Như đã trình bày trong mục 2.2.2, 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 [17], 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  $O(N_d \log(N_w))$ . 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 3.1, 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 3.1 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

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

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 3.2: Thí nghiệm lọc bỏ các stop words (các từ có tần số xuất hiện cao nhất trong bộ dữ liệu).

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.

#### 3.3.2 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 từ trực quan bằng cách đếm số lần xuất hiện của các từ trong các văn bản và lọc bỏ một nhóm các từ có số lần xuất hiện nhiều nhất. Kết quả thí nghiệm được tiến hành trên bộ Oxford 5k và được trình bày trong bảng 3.2.

Kết quả thí nghiệm trong bảng 3.2 cho thấy khi lọc bỏ stop word, độ chính xác của truy vấn tăng lên rõ rệt. Khi lọc bỏ 5% số từ 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 từ có tần suất xuất hiện cao nhất ở các thí nghiệm tiếp theo để hệ thống đạt được hiệu suất tốt nhất.

# 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 [17], 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) [26].

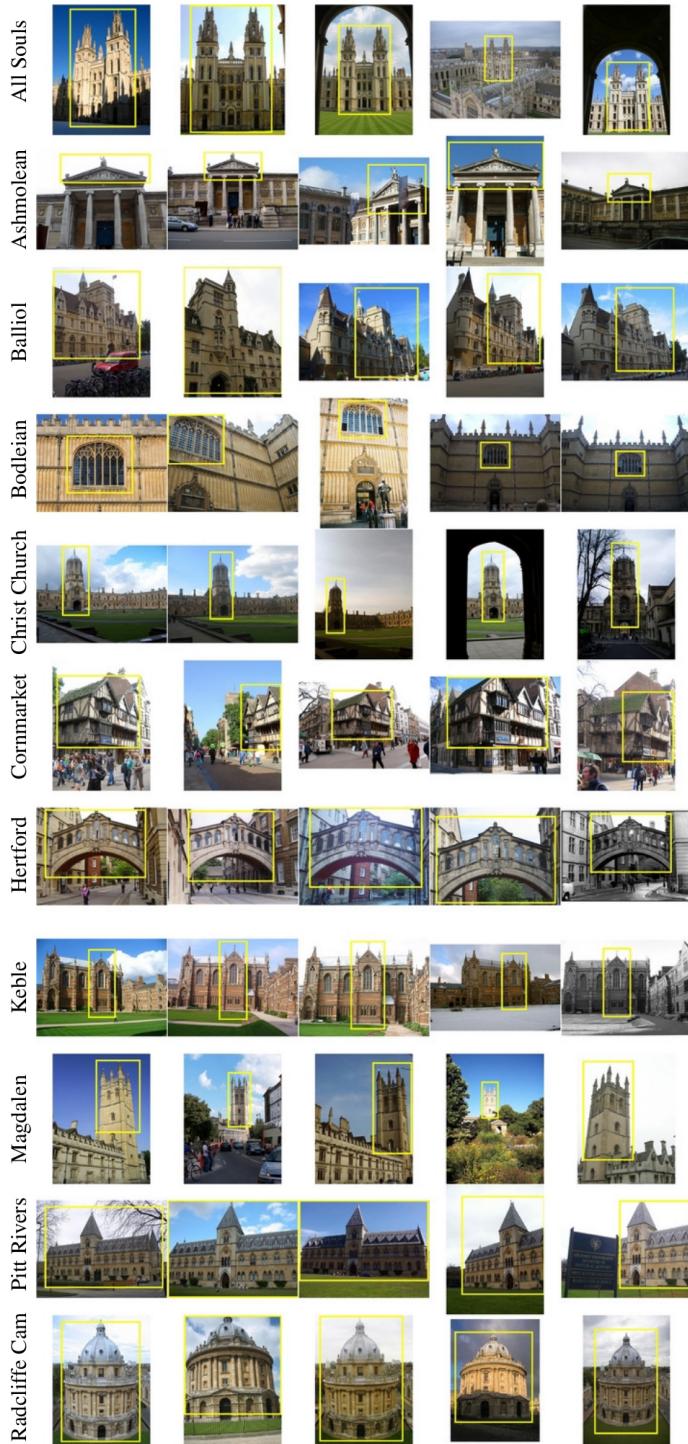
Paris 6k được đánh giá là một bộ dữ liệu hoàn toàn độc lập so với Oxford 5k và thường được dùng để kiểm tra các tác động của việc tính toán từ trực quan trong khi Oxford 5k thường được dùng để kiểm tra hiệu suất.

#### 4.1.1.3 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 [17].

## 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) [17], 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.

### 4.2 Cài đặt thí nghiệm

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 [24].
- **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 và các thông số cho phương pháp đề xuất.

**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 [28]. 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 [11] để đạt được hiệu suất tốt 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 [17], 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 từ trực quan.

**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

## 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 4.2: Hiệu suất của các phương pháp trên bộ dữ liệu Oxford 5k.

sẽ lưu thông tin cho một ô. Tất cả các tập chỉ mục đó sẽ được lưu thành 1 một 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 từ trực quan 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.

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

Bảng 4.2 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.

Trong khi đó, với việc tích hợp thông tin không gian ảnh vào 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

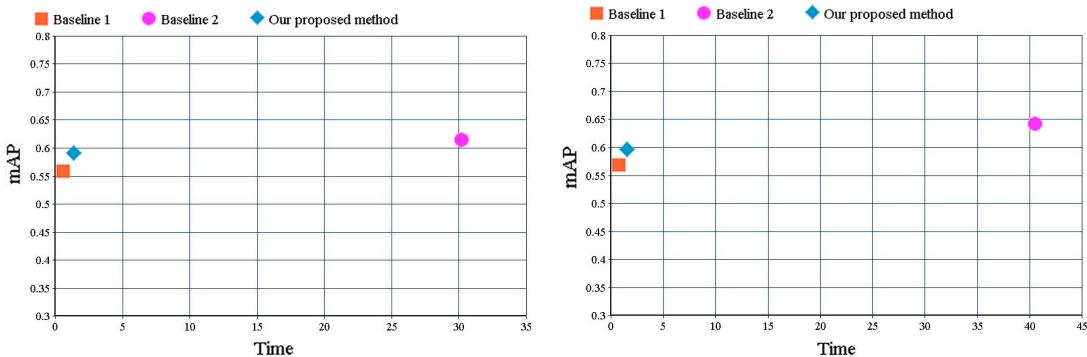
#### 4. Thực nghiệm và đánh giá kết quả

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 4.3: Hiệu suất của các phương pháp trên bộ dữ liệu Paris 6k.

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 4.3. Biểu đồ trong Hình 4.2 cho thấy sự so sánh tương quan giữa các phương pháp khi thử nghiệm trên bộ dữ liệu Oxford 5k.



Hình 4.2: 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).

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

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

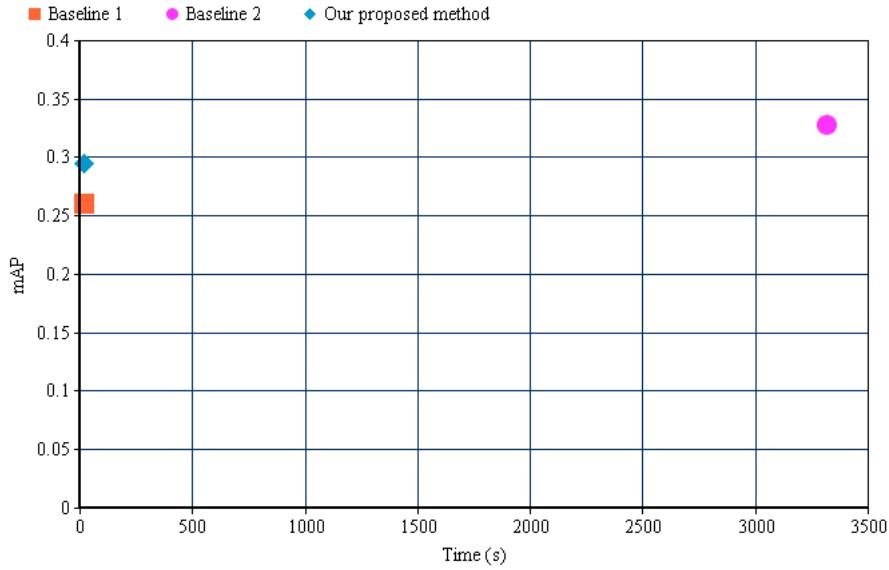
Phương pháp (trên Oxford 5K+100K)	mAP	Thời gian truy vấn (55 truy vấn)	Bộ nhớ sử dụng
Phương pháp cơ sở 1	0.2601	16.1319 (s)	364.14MB
Phương pháp cơ sở 2	0.3279	3315.02 (s)	364.14MB
<b>Phương pháp đề xuất (<math>L = 2</math>)</b>	<b>0.2950</b>	<b>18.4219 (s)</b>	<b>1,369.88MB</b>

Bảng 4.4: Hiệu suất của các phương pháp trên bộ dữ liệu Oxford 5K+100K.

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 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. Kết quả thí nghiệm trên bộ dữ liệu này được thể hiện trong Bảng 4.4. Có thể thấy rằng, mặc dù độ chính xác của các phương pháp đều giảm tuy nhiên 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 4.3 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 4.5 và Bảng 4.6. 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

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



Hình 4.3: Biểu đồ so sánh độ chính xác và thời gian truy vấn giữa các phương pháp trên bộ Oxford 5K+100K.

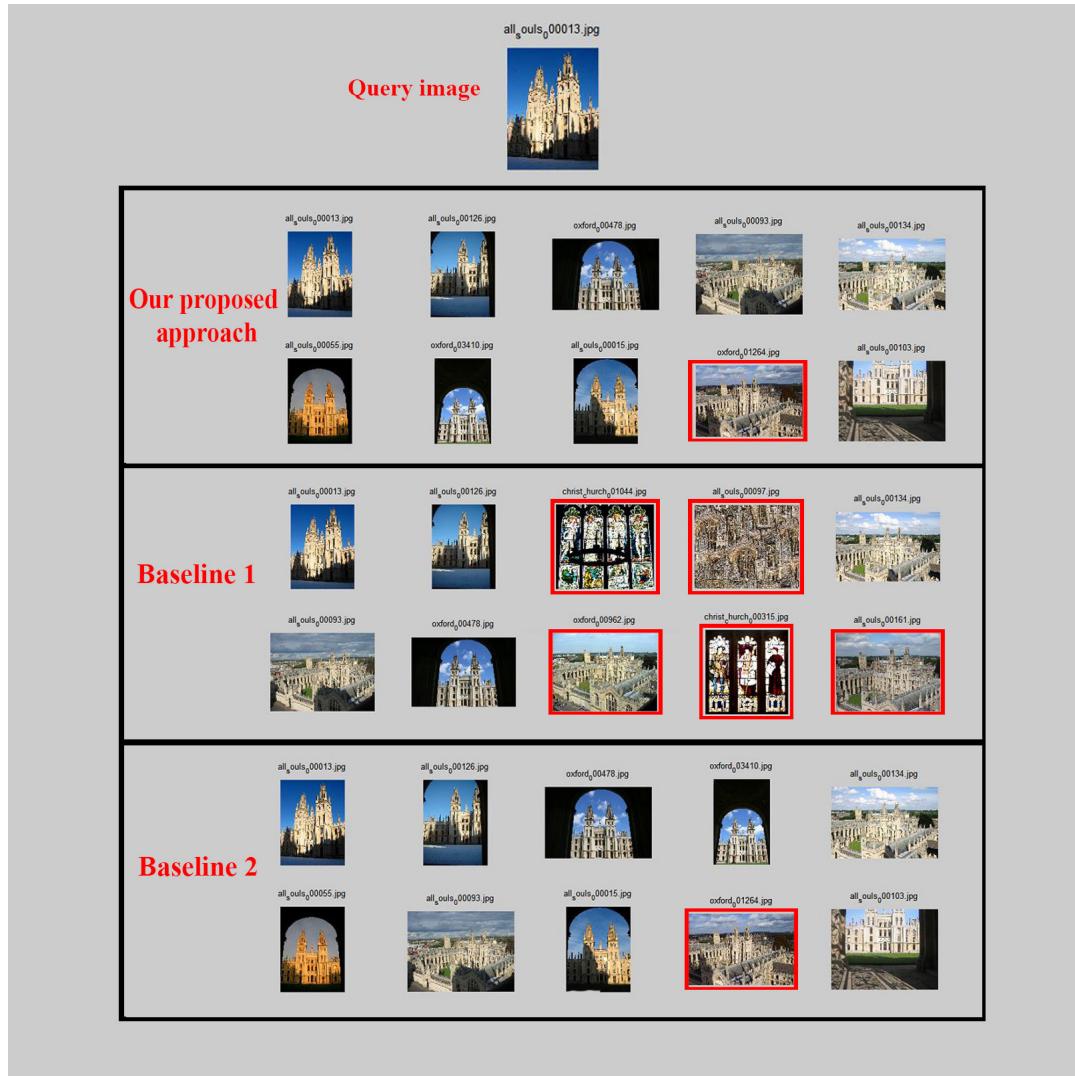
hiện rõ nét hơn. Tuy nhiên, không phải lúc nào  $L$  tăng thì độ chính xác cũng sẽ tăng theo.

Hình 4.4 cho thấy ví dụ về hình ảnh truy vấn và thẻ hiện các kết quả trả về với các phương pháp khác nhau trên bộ Oxford 5k. 10 hình ảnh có đứng 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.

$L$	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 4.5: Hiệu suất của phương pháp đề xuất với các giá trị khác nhau của  $L$  trên bộ dữ liệu Oxford 5K.

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



Hình 4.4: 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.

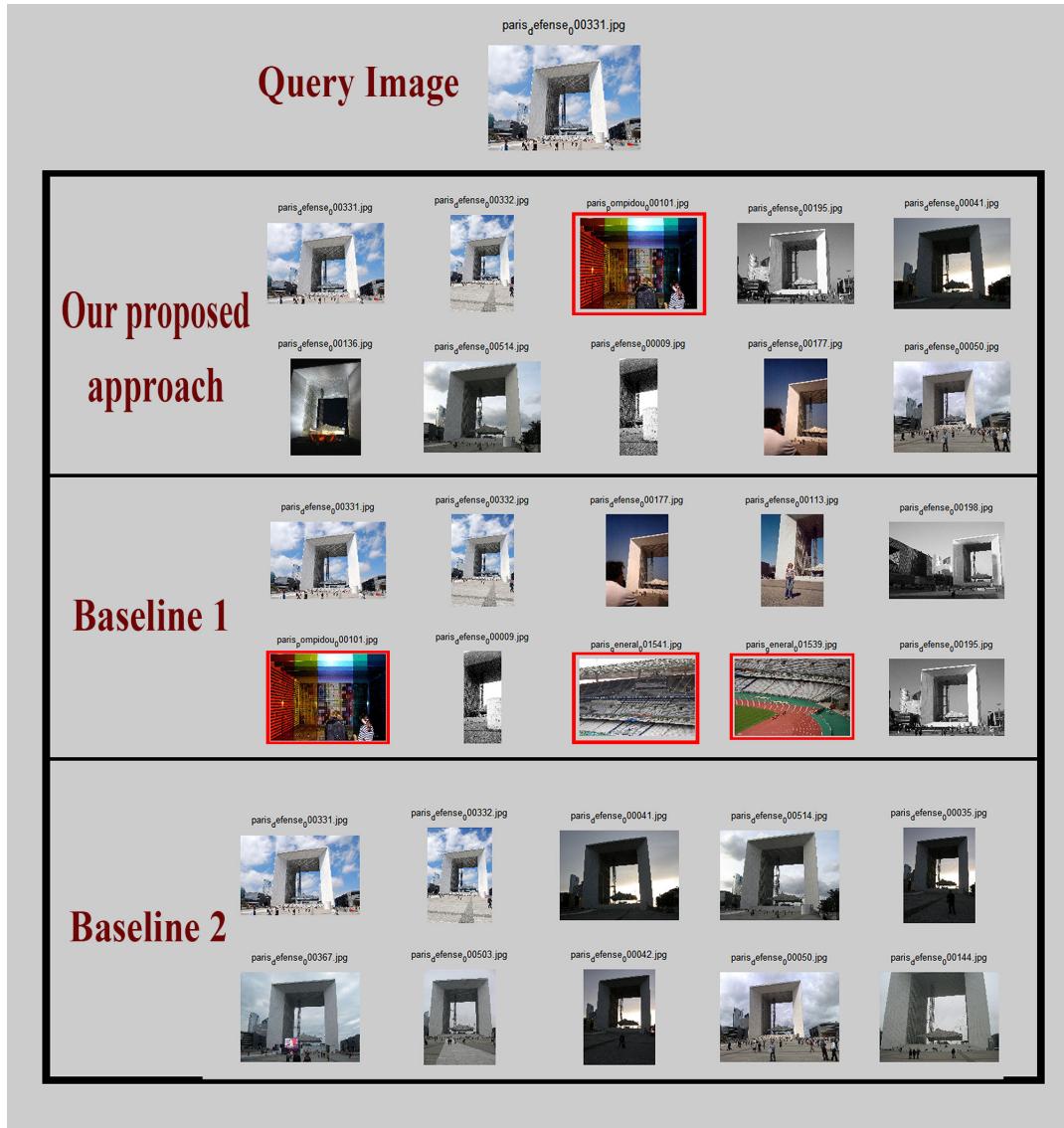
#### 4. Thực nghiệm và đánh giá kết quả

$L$	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 4.6: Hiệu suất của phương pháp đề xuất với các giá trị khác nhau của  $L$  trên bộ dữ liệu Paris 6K.

Tương tự, trong hình 4.5, 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



Hình 4.5: 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 và tăng tính ứng dụng cho 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 ..., 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**

---

### **5.1.2 Các chức năng chính**

Trong ứng dụng này, để tăng tính tiện dụng và gần gũi với người dùng, phía client được xây dựng trên nền tảng thiết bị động. Ứng dụng bao gồm các chức năng chính 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, thẻ nhớ để tìm kiếm.
- Chọn vùng đối tượng: từ ảnh chụp được hay ảnh được lưu trữ sẵn trong máy, để nâng cao độ chính xác của việc tìm kiếm, người dùng có thể khoanh vùng đối tượng cần tìm trên ảnh.

## **5.2 Thiết kế ứng dụng**

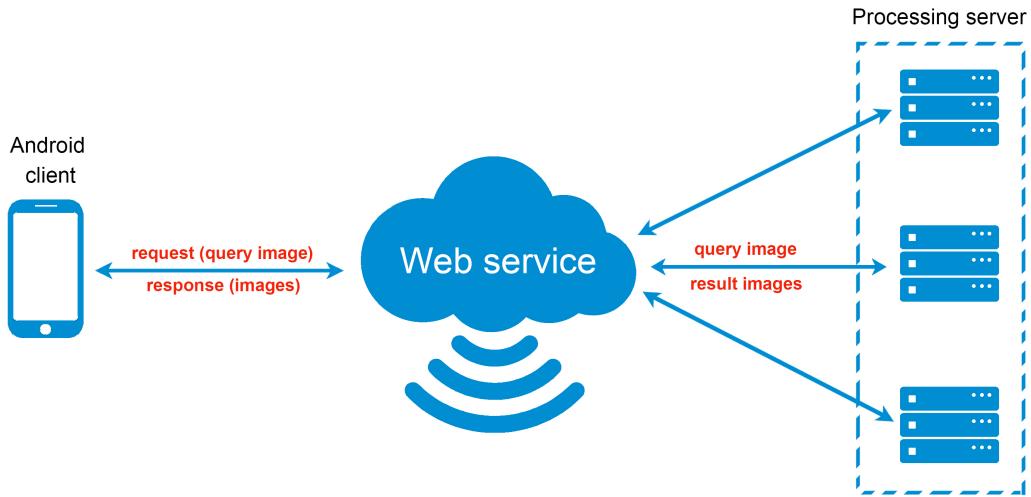
### **5.2.1 Kiến trúc**

Kiến trúc của chương trình được chia làm 3 thành phần chính:

- Client side: là một ứng dụng di động chạy trên hệ điều hành Android cho phép người dùng chụp ảnh hoặc chọn ảnh từ file, sau đó nén và mã hóa và gửi lên web service.
- Web service: nhận các request từ phía client. Thực hiện điều phối và cân bằng tải để chuyển tiếp các request cho các web server để xử lý. Đồng thời tiếp nhận kết quả xử lý từ các web server để trả về cho từng client tương ứng.
- Server side: Nhận request từ web service chuyển tiếp lên, xử lý và trả về kết quả cho web service. Việc xử lý ở phía server tốn rất nhiều tài nguyên thế nên nếu lượng request lớn, sẽ rất dễ rơi vào trường hợp thắt cổ chai. Do đó, để tăng sức mạnh cho hệ thống, ứng dụng có thể scale ra nhiều server để chia tải.

Hình [5.1](#) minh họa cho mô hình hoạt động tổng quan của hệ thống.

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



Hình 5.1: Kiến trúc tổng quan của hệ thống.

### 5.2.1.1 Client side

### 5.2.1.2 Web service

### 5.2.1.3 Server side

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 query.

Sơ đồ xử lý của quá trình training được thể hiện trong Hình 3.1. Như đã trình bày sơ lược 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 RootSIFT. Các vector thu được chính là các vector 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 AKM để thu được các visual word với thông số  $k = 1$  triệu. Kết quả ta sẽ thu được các visual word là các từ dùng để mô tả hình ảnh.

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

---

Từ những visual word này, ta sẽ xây dựng thành một từ điển để dễ dàng cho quá trình biểu diễn và 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.

Quá trình training được thực hiện độc lập trên server.

Quá trình truy vấn được đã trình bày chi tiết trong mục 3.2 và được thể hiện trong Hình 3.2. 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 áp dụng phương pháp xếp hạng voting để tìm được các ứng viên với số lượt bầu chọn nhiều nhất.

### **5.2.2 Giao diện**

## **5.3 Cài đặt và thử nghiệm**

### **5.3.1 Môi trường cài đặt**

### **5.3.2 Kết quả thử nghiệm**

### **5.3.3 Đánh giá kết quả**

# Chương 6

## Tổng kết

### 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 phương pháp nhằm nâng hiệu suất của các hệ thống truy vấn ảnh trên tập dữ liệu lớn phục vụ cho các ứng dụng yêu cầu 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 Holidays 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.

Mặc dù công trình nghiên cứu còn giới hạn và nhiều hạn chế song đã đạt được những thành quả bước đầu đáng khích lệ, làm nền tảng cho những nghiên cứu sau này.

### 6.2 Hướng phát triển

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 tỷ 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 từ trực quan hay không chứ không quan tâm tới con số của từ trực quan 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 từ giống nhau tập trung trong một ô cục bộ.

**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 từ 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 từ, số lượng của từ đó trong ảnh,...

# Tài liệu tham khảo

- [1] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004. [9](#), [10](#)
- [2] 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. [9](#)
- [3] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” *International journal of computer vision*, vol. 60, no. 1, pp. 63–86, 2004. [9](#)
- [4] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in *Computer Vision–ECCV 2006*. Springer, 2006, pp. 404–417. [9](#), [10](#)
- [5] 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. [9](#)
- [6] 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. [9](#)
- [7] 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. [10](#)

## TÀI LIỆU THAM KHẢO

---

- [8] 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. [10](#)
- [9] 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. [10](#)
- [10] 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. [10](#)
- [11] 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. [10, 29](#)
- [12] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval.* New York, NY, USA: McGraw-Hill, Inc., 1986. [11](#)
- [13] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval.* Cambridge university press Cambridge, 2008, vol. 1. [11, 12](#)
- [14] 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. [12, 13, 15](#)
- [15] ——, “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. [13](#)
- [16] 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. [13](#)
- [17] 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. [13](#), [15](#), [23](#), [26](#), [27](#), [29](#)

- [18] 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. [15](#)
- [19] 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. [15](#)
- [20] 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. [15](#)
- [21] 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. [15](#)
- [22] 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. [15](#)
- [23] 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. [15](#)
- [24] 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. [16](#), [19](#), [29](#)
- [25] K. Grauman and T. Darrell, “The pyramid match kernel: Discriminative classification with sets of image features,” in *Computer Vision, 2005. ICCV*

## TÀI LIỆU THAM KHẢO

---

2005. *Tenth IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1458–1465. [16](#)

- [26] 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. [26](#)
- [27] H. Jégou, M. Douze, and C. Schmid, “Hamming embedding and weak geometric consistency for large scale image search,” in *European Conference on Computer Vision*, ser. LNCS, A. Z. David Forsyth, Philip Torr, Ed., vol. I. Springer, oct 2008, pp. 304–317. [Online]. Available: <http://lear.inrialpes.fr/pubs/2008/JDS08>
- [28] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, 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. [29](#)