MỤC LỤC

DANH MỤC HÌNH VẼ

DANH SÁCH CÁC TỪ VIẾT TẮT

STT	Từ viết tắt	Mô tả
1	APC	Affinity Propagation Clustering
2	ARE	Augmented Relation Embedding
3	CBIR	Content-Based Image Retrieval
4	CNS	Color Naming System
5	CRT	Composite Region Template
6	FSVM	Fuzzy Support Vector Machine
7	HSL	Hue-Saturation-Luminance
8	KL	Karhunen-Loeve
9	LGRM	Local and Global Regressive Mapping
10	LPC	Locality Preserving Clustering
11	MRBIR	Manifold-Ranking Based Image Retrieval
12	MSRA	Microsoft Research Asia
13	NCut	Normalized Cut
14	PCA	Principal Component Analysis
15	RF	Relevance Feedback
16	RGB	Red-Green-Blue
17	ST	Semantic Template
18	SVM	Support Vector Machine
19	SVT	Semantic Visual Template

LỜI CẢM ƠN

Em xin chân thành cảm ơn Thầy giáo, Thạc sĩ Ngô Trường Giang, người đã hướng dẫn tận tình chỉ bảo em rất nhiều trong suốt quá trình tìm hiểu nghiên cứu và hoàn thành đồ án này từ lý thuyết đến ứng dụng. Sự hướng dẫn của thầy đã giúp em có thêm kiến thức về lập trình và kiến thức về lĩnh vực xử lý ảnh.

Đồng thời, em xin chân thành cám ơn các thầy cô trong khoa Công nghệ thông tin – Trường Đại Học Dân Lập Hải Phòng, cũng như các thầy cô trong trường đã trang bị cho em những kiến thức cơ bản cần thiết trong suốt thời gian học tập tại trường để em hoàn thành tốt đồ án này.

Em xin chân thành cảm ơn GS.TS.NGƯT Trần Hữu Nghị, Hiệu trưởng Trường Đại học Dân Lập Hải Phòng, ban giám hiệu nhà trường, khoa Công nghệ thông tin, các phòng ban nhà trường đã tạo điều kiện tốt nhất trong suốt thời gian em học tập và làm tốt nghiệp.

Trong quá trình học cũng như trong suốt thời gian làm đồ án tốt nghiệp không tránh khỏi những thiếu sót, em rất mong được sự góp ý quý báu của các thầy cô cũng như tất cả các bạn để kết quả của em được hoàn thiện hơn.

Sau cùng, em xin gửi lời cảm ơn đến gia đình, bạn bè đã tạo mọi điều kiện để em xây dựng thành công đồ án này.

Em xin chân thành cảm ơn!

Hải Phòng, ngày 2 tháng 12 năm 2014

Sinh viên thực hiện

Trương Thanh Tùng

MỞ ĐẦU

Từ khi mạng internet được toàn cầu hoá kéo theo sự mở rộng của các nội dung đa phương tiện như: ảnh, âm nhạc, video, v.v…, khiến cho nhu cầu chia sẻ và tìm kiếm các loại tài nguyên này cũng tăng theo một cách nhanh chóng. Và trong số các tài nguyên đó phải kể đến các dữ liệu hình ảnh. Mỗi người đều tìm kiếm hình ảnh theo một mục đích khác nhau, nhưng chung quy lại cái người dùng muốn tìm kiếm là những thông tin kèm theo và hình ảnh mà họ cần hoặc một số khác lại tìm kiếm hình ảnh để xác nhận tính chính xác của thông tin họ nhận được. Các dữ liệu hình ảnh đều phục vụ cho nhiều lĩnh vực quan trọng trong cuộc sống như trong các hệ thống bảo mật, an ninh, y tế, hay các hệ thống phát hiện chuyển động …. Vì thế việc nghiên cứu và phát triển các hệ thống tra cứu ảnh ngày càng trở nên cấp thiết.

Đồ án sẽ trình bày kỹ thuật phản hồi liên quan được ứng dụng trong tra cứu ảnh dựa trên nội dung để thu hẹp khoảng cách ngữ nghĩa, cải thiện hiệu năng tra cứu. Báo cáo được chia làm 3 chương:

Chương 1: Các khái niệm cơ bản về tra cứu ảnh dựa trên nội dung

Giới thiệu các khái niệm cơ bản về hệ thống tra cứu ảnh dựa trên nội dung và các thành phần trong nó.

Chương 2: Tra cứu ảnh dựa trên nội dung với phản hồi liên quan

Giới thiệu các phương pháp phản hồi liên quan và máy học SVM.

Chương 3: Áp dụng cài đặt thử nghiệm

Cài đặt chương trình thử nghiệm dựa trên lý thuyết từ chương 1 và chương 2.

CHƯƠNG 1: CÁC KHÁI NIỆM CƠ BẢN VỀ TRA CỨU ẢNH DỰA TRÊN NỘI DUNG

1.1 Giới thiệu

Trong thời đại bùng nổ về tìm kiếm thông tin hiện nay, ngoài việc tìm kiếm các văn bản nội dung thì việc tìm kiếm ảnh đang có xu hướng trở nên phổ biến. Với nguồn tài nguyên ảnh vô cùng to lớn trên mạng internet, thì việc tìm kiếm chính xác một bức ảnh đúng với yêu cầu của người dùng là khó khả thi. Chúng ta khó có thể tìm kiếm một bức ảnh theo cách thông thường, có nghĩa là việc tìm kiếm được thực hiện lần lượt trên từng tấm ảnh cho đến khi tìm thấy đúng ảnh có nội dung cần tìm. Với lại nguồn tài nguyên ảnh trên mạng internet sẽ ngày càng nhiều hơn nữa theo sự phát triển của công nghệ số trong tương lai. Do đó, nhu cầu thật sự đòi hỏi chúng ta phải có một công cụ hỗ trợ cho việc tìm kiếm ảnh càng sớm càng tốt.

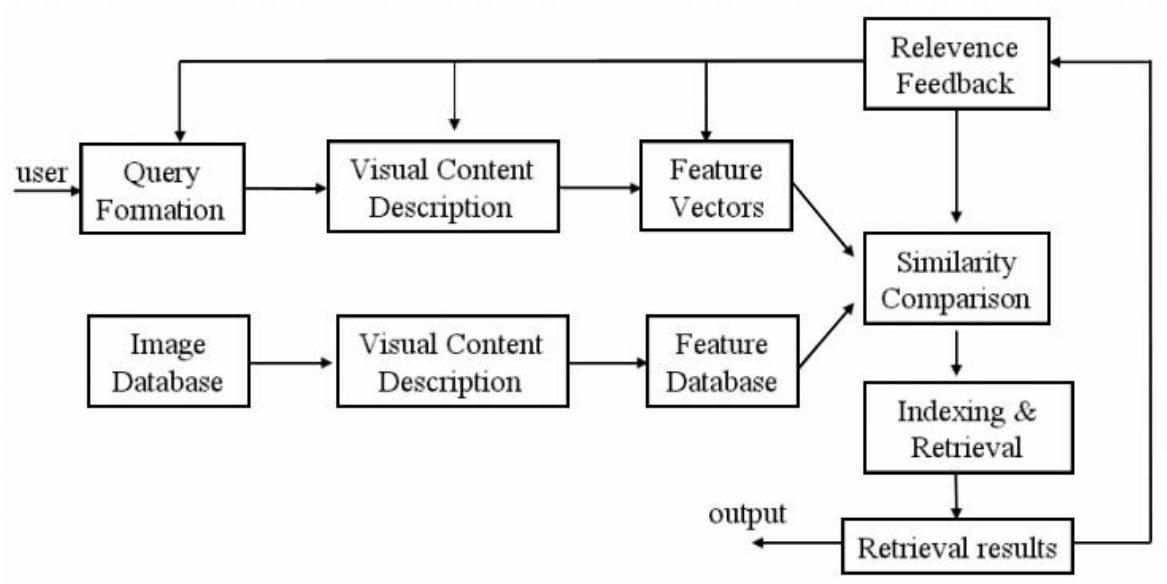
Có hai kiểu tìm kiếm dữ liệu ảnh đó là tìm kiếm theo từ khoá và tìm kiếm theo nội dung ảnh. Tìm kiếm theo từ khoá dễ thoả mãn được nhu cầu người dùng với các nhu cầu tìm kiếm hình ảnh mới theo mong muốn bản thân họ. Và thêm nữa, tìm kiếm theo từ khoá thì nhanh hơn tìm kiếm theo nội dung bởi vì nó hoạt động trên việc phân tích và so sánh các từ hoặc cụm từ tương ứng với nhau để đưa ra kết quả. Kiểu dữ liệu này là dạng các văn bản, từ ngữ cho nên kết quả được đưa ra nhanh chóng, và không đòi hỏi người dùng phải có ảnh mẫu. Tuy nhiên, phương pháp này có nhược điểm là kết quả hình ảnh không phải lúc nào cũng chính xác so với mong muốn của người dùng. Và nó chỉ phù hợp nhất với việc đáp ứng nhu cầu của người dùng thông qua các mô tả bằng từ ngữ. Để khắc phục nhược điểm này của tìm kiếm theo từ khoá, một phương pháp khác được ra đời là tìm kiếm ảnh theo nội dung. Ý tưởng chính của phương pháp này là tạo ra sự mô tả ảnh một cách tự động trực tiếp từ nội dung ảnh bằng sự phân tích nội dung ảnh mà không có sự can thiệp thủ công. Ưu điểm của phương pháp này so với phương pháp dựa trên từ khóa là việc trích chọn đặc trưng được thực hiện một cách tự động và nội dung của ảnh luôn luôn nhất quán. Trong thực tế, con người có xu hướng sử dụng các đặc trưng mức cao (khái niệm), chẳng hạn như từ khóa, mô tả văn bản để giải thích hình ảnh và đo độ tương tự của chúng. Trong khi những đặc trưng được trích chọn một cách tự động bằng cách sử dụng kỹ thuật của thị giác máy chủ yếu là các đặc trưng mức thấp (màu sắc, kết cấu, hình dạng, vị trí không gian…). Mặc dù nhiều thuật toán phức tạp đã được thiết kế để mô tả màu sắc, hình dáng và đặc trưng kết cấu nhưng các thuật toán này vẫn không thể phản ánh thỏa đáng ngữ nghĩa ảnh.

Do vậy, khoảng cách ngữ nghĩa giữa các đặc trưng mức thấp và các khái niệm mức cao vẫn còn lớn nên hiệu suất của *CBIR* là vẫn còn xa với mong đợi của người dùng.

Để thu hẹp khoảng cách ngữ nghĩa, phản hồi liên quan (*RF*) được xem như là một công cụ hiệu quả để cải thiện hiệu năng của hệ thống *CBIR*. Nói chung, *RF* nhằm mục đích cải thiện hiệu năng tra cứu thông qua việc học những điều chỉnh của người dùng trên những kết quả tra cứu. Theo cách này, hệ thống cần phải thực hiện thông qua một số vòng lặp. Trong mỗi vòng lặp, hệ thống sẽ trả lại một danh sách ngắn các ảnh tương tự nhất với ảnh truy vấn dựa trên khoảng cách *Euclidean*. Sau đó, các ảnh này sẽ được đưa cho người sử dụng gán nhãn liên quan hoặc không liên quan với ảnh truy vấn. Sử dụng những ảnh đã được gán nhãn này như là những hạt giống, những kỹ thuật học máy sẽ được áp dụng để xây dựng mô hình phân lớp tất cả các ảnh trong cơ sở dữ liệu thành hai lớp liên quan và không liên quan với ảnh truy vấn. Hàm phân lớp sau đó được sử dụng như hàm xếp hạng để đo độ liên quan của ảnh trong cơ sở dữ liệu.

1.2 Những thành phần của một hệ thống tra cứu ảnh dựa trên nội dung

Một hệ thống tra cứu ảnh đòi hỏi các thành phần như trong hình 1-1 [5].



Hình 1-1: Kiến trúc tổng quan về hệ thống tra cứu ảnh

Trong đó có ba thành phần quan trọng nhất trong tra cứu ảnh dựa trên nội dung: Trích chọn đặc trưng, đánh chỉ số và giao diện truy vấn cho người dùng.

1.2.1 Trích chọn đặc trưng (Features Extraction):

Các đặc trưng của ảnh bao gồm các đặc tính cơ bản và các đặc tính ngữ nghĩa/logic. Các đặc tính cơ bản đó là: màu sắc (*color*), hình dạng (*shape*), kết cấu (*texture*), vị trí không gian (*spatial location*). Chúng có thể được trích xuất tự động hoặc bán tự động. Đặc tính *logic* cung cấp mô tả trừu tượng của dữ liệu hình ảnh ở các cấp độ khác nhau. Thông thường, các đặc tính *logic* được chiết xuất bằng tay hoặc bán tự động. Một hoặc nhiều đặc trưng có thể được sử dụng trong ứng dụng cụ thể.

1.2.1.1 Đặc trưng màu sắc (color):

Đặc trưng màu sắc là một trong những đặc tính được sử dụng phổ biến trong tra cứu ảnh. Màu sắc được định nghĩa trên một không gian màu lựa chọn. Sự đa dạng của không gian màu là có sẵn, chúng thường được dùng cho các ứng dụng khác nhau. Không gian màu được thể hiện gần gũi hơn với nhận thức của con người và được sử dụng rộng rãi trong *RGB, LAB, HSV, ...* Đặc trưng màu sắc phổ biến hoặc các mô tả trong hệ thống *CBIR* bao gồm: ma trận hiệp biến màu, biểu đồ màu, *moment* màu, và *véc-tơ* kết hợp màu [5]. Vào năm 1999, Gevers và cộng sự đã quan tâm đến các đối tượng lấy từ các điểm quan sát khác nhau và sự chiếu sáng. Theo kết quả, một tập các điểm bất biến đặc trưng màu đã được tính toán. Các bất biến màu được xây dựng trên cơ sở của "*hue*", "*cặp hue-hue*", và ba đặc trưng màu được tính toán từ các mô hình đối xứng. Việc lựa chọn đặc trưng màu phụ thuộc vào kết quả phân đoạn. Ví dụ, nếu việc phân đoạn cung cấp đối tượng mà không có màu đồng nhất, thì rõ ràng màu trung bình không phải là lựa chọn tốt. Với các ứng dụng đặc biệt như cơ sở dữ liệu khuôn mặt người, thì miền tri thức có thể được khai thác để gán trọng số cho từng điểm ảnh trong việc tính toán vùng màu.

1.2.1.2 Đặc trưng kết cấu (texture):

Kết cấu không được định nghĩa đầy đủ như là đặc trưng màu sắc, vì thế mà một số hệ thống không sử dụng đặc trưng kết cấu. Tuy nhiên, kết cấu cung cấp các thông tin quan trọng trong việc phân loại ảnh, vì nó mô tả nội dung của nhiều ảnh thực như là: vỏ trái cây, mây, cây, gạch, và vải. Do đó, kết cấu là một tính năng quan trọng trong việc định nghĩa ngữ nghĩa mức cao cho mục đích tra cứu ảnh [5]. Các đặc trưng kết cấu thường được sử dụng trong hệ thống tra cứu ảnh bao gồm các đặc trưng phổ, chẳng hạn như các đặc trưng được bao gồm sử dụng lọc *Gabor* hoặc biến đổi *wavelet*, thống kê đặc trưng kết cấu trong các cách đo độ thống kê cục bộ, như sáu đặc trưng kết cấu *Tamura*, và đặc trưng *wold* được đề xuất bởi Liu và các cộng sự vào năm 1996.

1.2.1.3 Đặc trưng dựa trên hình dạng (shape):

Hình dạng là một khái niệm được định nghĩa khá tốt. Đặc trưng hình dạng của các ứng dụng nói chung bao gồm: tỷ lệ *aspect*, tuần hoàn, mô tả *Fourier*, bất biến thời điểm, phân đoạn đường bao liên tiếp, …. Đặc trưng hình dạng là đặc trưng ảnh quan trọng, mặc dù chúng chưa được sử dụng rộng rãi trong *CBIR* như là đặc trưng màu và đặc trưng kết cấu [5]. Đặc trưng hình dạng đã thể hiện tính hữu ích trong nhiều miền ảnh đặc biệt như là các đối tượng nhân tạo. Đối với ảnh màu thì được sử dụng trong hầu hết các loại giấy tờ, tuy nhiên, nó lại khó khăn để áp dụng đặc trưng hình dạng so với màu sắc và kết cấu do sự thiếu chính xác của phân đoạn. Mặc dù gặp khó khăn, đặc trưng hình dạng chỉ được sử dụng trong một số hệ thống và cho thấy tiềm năng có ích cho *CBIR*. Ví dụ, vào năm 2003, Mezaris và các cộng sự đã sử dụng các đặc trưng hình dạng đơn giản chẳng hạn như: độ lệch tâm và định hướng. Một hệ thống mà Wang và cộng sự vào năm 1999 đã sử dụng tiêu chuẩn hoá quán tính của thứ tự từ 1-3 để mô tả hình dạng khu vực.

1.2.1.4 Đặc trưng không gian (spatial location):

Các vùng hoặc đối tượng với thuộc tính màu sắc và kết cấu tương tự có thể được nhận ra một cách dễ dàng bởi ràng buộc không gian [5]. Ví dụ, các vùng có bầu trời và biển màu xanh có thể có biểu đồ màu tương tự, nhưng lại có vị trí không gian trong ảnh khác nhau. Vì thế, vị trí không gian của các vùng (hoặc các đối tượng) hoặc mối liên hệ không gian giữa nhiều vùng trong một ảnh thì rất hữu dụng cho việc tra cứu ảnh. Một biểu diễn của mối liên hệ không gian được sử dụng rộng rãi nhất là "*2D strings*" được Chang và các cộng sự đưa ra vào năm 1987. Kỹ thuật này được xây dựng bằng cách chiếu các ảnh theo trục *x* và *y*. Cho hai tập ký hiệu *V* và *A*, được định nghĩa trên phép chiếu. Cứ mỗi ký hiệu trong *V* thì được biểu diễn bởi một đối tượng trong ảnh. Còn mỗi ký hiệu trong *A* thì được biểu diễn bởi một loại liên hệ không gian giữa các đối tượng. Nếu chúng khác nhau, thì kỹ thuật "*2D G-string*" sẽ cắt tất cả các đối tượng dọc theo đường bao hộp nhỏ nhất và mở rộng mối liên hệ không gian vào trong hai tập toán tử không gian. Một tập toán tử thì định nghĩa mối liên hệ không gian cục bộ. Và tập còn lại thì định nghĩa mối liên hệ không gian toàn cục, chỉ ra rằng phép chiếu của hai đối tượng là tách ra, nối liền hoặc được xác định ở cùng vị trí. Ngoài ra, kỹ thuật "*2D C-string*" thì được đề xuất bởi Lee và các cộng sự vào năm 1990 để cực tiểu con số đối tượng đã cắt. Còn kỹ thuật "*2D B-string*" thì lại được giới thiệu vào năm 1992 bởi Yang và các cộng sự. Kỹ thuật này sẽ biểu diễn một đối tượng bởi hai ký hiệu, thay thế cho việc mở đầu và kết thúc đường bao của đối tượng.

Hầu hết các phương pháp trên có thể tạo ra ba kiểu truy vấn. Kiểu truy vấn 0 sẽ tìm tất cả các ảnh chứa đối tượng O_1, O_2, ..., O_n. Kiểu truy vấn 1 sẽ tìm tất cả các ảnh chứa đối tượng mà có mối liên hệ chắc chắn giữa từng đối tượng khác, nhưng khoảng cách giữa chúng là không có nghĩa. Cuối cùng, kiểu truy vấn 2 sẽ tìm tất cả các ảnh mà có liên kết khoảng cách chắc chắn với từng đối tượng khác.

Ngoài kỹ thuật "*2D string*" ở trên, còn có các kỹ thuật khác như: "*spatial quad-tree*" được giới thiệu vào năm 1984 bởi Samet, và "*symbolic image*" được giới thiệu vào năm 1995 bởi Gudivada và Raghavan. Hai kỹ thuật này thì được dùng để biểu diễn thông tin không gian. Tuy nhiên, tra cứu ảnh dựa trên mối liên hệ không gian của vùng thì vẫn là một bài toán khó trong nghiên cứu tra cứu ảnh dựa trên nội dung. Bởi vì, các phân đoạn của đối tượng hoặc các vùng đáng tin cậy thì thường không khả thi trừ khi trong các ứng dụng rất giới hạn. Mặc dù, một số hệ thống đơn giản phân chia ảnh vào trong các khối con chuẩn, mà chỉ đạt được thành công nhỏ với phương pháp phân chia không gian từ hầu hết ảnh tự nhiên không bị nén vào các khối con chuẩn. Để giải quyết vấn đề này, một phương pháp dựa trên kỹ thuật "*Radon transform*", một kỹ thuật mà sẽ khai thác các đặc trưng trực quan của sắp xếp không gian mà không cần phân đoạn phức tạp. Phương pháp này được đề xuất vào năm 1998 bởi Guo và các cộng sự.

1.2.2 Đánh chỉ số (Indexing):

Một vấn đề quan trọng khác trong tra cứu ảnh dựa trên nội dung là đánh chỉ số và tìm kiếm nhanh ảnh dựa trên đặc trưng trực quan. Bởi vì, các *véc-tơ* đặc trưng của ảnh có xu hướng tới số chiều cao và do đó nó không thích hợp cho các cấu trúc đánh chỉ số truyền thống. Việc giảm số chiều thường xuyên được sử dụng trước khi lên kế hoạch đánh chỉ số.

Một trong những công nghệ được sử dụng phổ biến cho việc giảm số chiều là Phân tích thành phần chính *PCA* [5]. Nó là một công nghệ tối ưu trong việc ánh xạ tuyến tính dữ liệu đầu vào cho một không gian toạ độ. Các trục được thẳng hàng để ánh xạ tối đa các biến in dữ liệu. Hệ thống *QBIC* sử dụng *PCA* để làm giảm 20 chiều trong véc-tơ đặc trưng hình dạng thành hai hoặc ba chiều. Ngoài công nghệ *PCA* ra, nhiều nhà nghiên cứu còn sử dụng biến đổi *KL* để làm giảm số chiều trong không gian đặc trưng. Mặc dù, biến đổi *KL* có một số thuộc tính hữu dụng như khả năng xác định vị trí hầu hết không gian con quan trọng, các thuộc tính đặc trưng mà quan trọng đối với việc xác định mô hình tương tự có thể bị phá huỷ trong suốt quá trình giảm các chiều mù. Ngoài hai công nghệ biến đổi *PCA* và *KL*, thì mạng *nơ-ron* cũng được chứng minh là công cụ hữu ích cho việc giảm số chiều đặc trưng.

1.2.3 Giao diện truy vấn (Query Interface):

Để biểu diễn ảnh tra cứu từ cơ sở dữ liệu cho người dùng thì có rất nhiều cách. Và những cách thông thường nhất được sử dụng là: duyệt qua mục, truy vấn bởi khái niệm, truy vấn bởi bản phác thảo, và truy vấn bởi ví dụ. Duyệt qua mục là phương pháp duyệt qua toàn bộ cơ sở dữ liệu theo danh mục các ảnh. Mục đích của phương pháp này là ảnh trong cơ sở dữ liệu được phân loại thành nhiều mục khác nhau theo ngữ nghĩa hoặc nội dung trực quan. Truy vấn bởi khái niệm là tra cứu ảnh theo mô tả khái niệm liên quan với từng ảnh trong cơ sở dữ liệu [5]. Truy vấn bởi bản phác thảo và truy vấn bởi ví dụ là vẽ ra một bản phác thảo hoặc cung cấp một ảnh ví dụ từ những ảnh với độ tương tự đặc trưng trực quan sẽ được trích chọn từ cơ sở dữ liệu.

Trong số ba tác vụ trên, thì trích chọn đặc trưng (*bao gồm cả đo độ tương tự*) là nhiệm vụ quan trọng và khó khăn nhất. Phần lớn các nghiên cứu tra cứu ảnh dựa trên nội dung tập trung đi sâu vào nhiệm vụ này.

1.3 Khoảng cách ngữ nghĩa trong tra cứu ảnh dựa trên nội dung

Trong lĩnh vực tra cứu ảnh hiện nay có hai hệ thống đang được phát triển là: hệ thống tra cứu dựa trên nội dung và hệ thống dựa trên từ khoá. Điểm khác biệt duy nhất giữa hai hệ thống này chính là sự tương tác của người dùng. Con người thì luôn có xu hướng sử dụng các khái niệm đặc trưng mức cao như là: từ khoá, mô tả văn bản, giải thích hình ảnh và đo độ tương tự. Trong khi đó, các đặc trưng ảnh được tự động trích chọn bằng kỹ thuật thị giác máy tính thì chủ yếu là các đặc trưng mức thấp (*màu sắc, kết cấu, hình dạng, vị trí không gian, v.v…*). Nói chung là không có mối liên quan trực tiếp giữa đặc trưng mức thấp và đặc trưng mức cao.

Mặc dù, các nhà nghiên cứu đã phát triển rất nhiều các thuật toán phức tạp để mô tả các đặc trưng: màu sắc, kết cấu, hình dạng. Thế nhưng, các thuật toán đó cũng không thể mô tả đầy đủ ngữ nghĩa của hình ảnh, và có nhiều hạn chế khi giải quyết một cơ sở dữ liệu nội dung ảnh lớn. Các thí nghiệm mở rộng trên hệ thống CBIR cho thấy nội dung đặc trưng mức thấp thường không thể mô tả các khái niệm ngữ nghĩa mức cao trong suy nghĩ người dùng. Do đó, hiệu suất của CBIR vẫn chưa đáp ứng được nhu cầu của người dùng. Tác giả Eakins vào năm 1999 đã đề xuất ra ba mức độ của các truy vấn trong CBIR.

- Mức 1: Tra cứu bởi các đặc trưng cơ bản như: màu sắc, kết cấu, hình dạng hoặc bố trí không gian của các phần tử ảnh.

- Mức 2: Tra cứu bởi các đối tượng được xác định bằng đặc trưng nguyên thuỷ, với một mức độ suy luận logic. Ví dụ: "tìm một bức ảnh có chứa bông hoa màu đỏ"
- Mức 3: Tra cứu bởi các thuộc tính trừu tượng, bao hàm số lượng mục đích các đối tượng trong ảnh, hoặc nội dung của ảnh được miêu tả. Điều này có nghĩa là tra cứu tên các sự kiện, ý nghĩa của ảnh, hoặc các dấu hiệu nổi bật, … Ví dụ như: "tìm một bức ảnh có đám đông vui vẻ".

Có thể thấy mức 2 và mức 3 được gọi là tra cứu ảnh ngữ nghĩa, và khoảng cách giữa mức 1 và mức 2 là khoảng cách ngữ nghĩa. Sự khác biệt giữa giới hạn mô tả đặc trưng ảnh mức thấp và sự phong phú ngữ nghĩa người dùng, được gọi là "*Khoảng cách ngữ nghĩa*".

1.4 Các phương pháp làm giảm khoảng cách ngữ nghĩa

Làm thế nào để chúng ta có thể liên kết các đặc trưng mức thấp của ảnh với các ngữ nghĩa mức cao?. Câu hỏi này đã thúc đẩy các nhà nghiên cứu cố gắng phát triển các công nghệ để giải quyết vấn đề này. Các công nghệ mới trong việc làm giảm khoảng cách ngữ nghĩa hiện nay có thể được phân ra theo các tiêu chí khác nhau. Bằng cách áp dụng vào các lĩnh vực khác nhau, các công nghệ tra cứu ảnh có thể có thể được chia ra là: tra cứu ảnh nghệ thuật, tra cứu ảnh phong cảnh, tra cứu ảnh web, v.v…. Dưới đây là một số kỹ thuật thường được sử dụng để suy ra ngữ nghĩa mức cao:

(1) Sử dụng bản thể đối tượng để định nghĩa khái niệm mức cao.

(2) Sử dụng phương pháp học có giám sát hoặc không có giám sát để gắn đặc trưng mức thấp với các khái niệm truy vấn.

(3) Giới thiệu phản hồi liên quan (*RF*) vào vòng lặp tra cứu ảnh cho việc học liên tục ý định của người dùng.

(4) Sinh mẫu ngữ nghĩa (*ST*) để hỗ trợ tra cứu ảnh mức cao.

(5) Sử dụng cả hai cách là thông tin văn bản từ trên web và nội dung trực quan của ảnh để tra cứu ảnh web.

Trong tất cả các kỹ thuật trên thì kỹ thuật (3) có thể rất khó được áp dụng và ít được phổ biến rộng rãi. Vì thế mà kỹ thuật (3) chỉ có thể được tìm thấy ở những miền lĩnh vực đặc biệt như là: bảo tàng nghệ thuật hay là các thư viện báo chí. Và hiện nay chỉ có kỹ thuật (2) là được áp dụng rộng rãi trong đời sống. Và những hệ thống áp dụng kỹ thuật (2) thì có 3 thành phần cơ bản như sau:

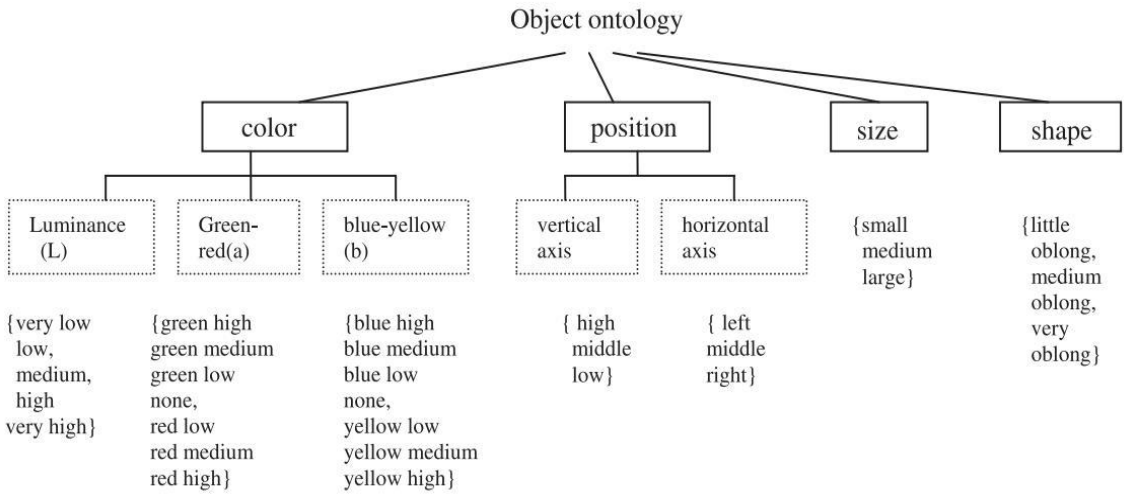
- Trích chọn đặc trưng ảnh mức thấp.
- Đo độ tương tự.
- Làm giảm khoảng cách ngữ nghĩa.

Thêm nữa là, có một số hệ thống chỉ sử dụng một kỹ thuật đã nêu trên để tra cứu ảnh dựa trên ngữ nghĩa mức cao. Nhưng cũng có một số hệ thống sử dụng kết hợp 3 hoặc 4 kỹ thuật ở trên.

1.4.1 Kỹ thuật bản thể đối tượng

Trong một số trường hợp, ngữ nghĩa có thể dễ dàng được suy ra từ ngôn ngữ hằng ngày. Ví dụ: "*bầu trời*" có thể được mô tả như là: "*ở trên, đồng đều, màu xanh da trời*". Trong các hệ thống sử dụng ngữ nghĩa đơn giản, trước tiên, các khoảng cách khác nhau được định nghĩa cho các đặc trưng ảnh mức thấp. Với mỗi khoảng cách tương ứng sẽ được mô tả mức trung gian của ảnh, ví dụ như: "*xanh nhạt, xanh vừa, xanh đậm*". Những mô tả đó sẽ hình thành một mẫu từ vựng đơn giản, và cái đó chúng ta gọi là "*bản thể đối tượng*". Đó là cái mà cung cấp định nghĩa đặc trưng của khái niệm truy vấn mức cao.

Cơ sở dữ liệu ảnh có thể được phân loại vào các mục khác nhau bằng cách ánh xạ mô tả ngữ nghĩa mức cao (*các từ khoá*) dựa trên kiến thức của con người. Ví dụ như: "*bầu trời*" có thể được định nghĩa như sau: "*xanh nhạt*" (*màu sắc*), "*đồng đều*" (*kết cấu*), "*trên cao*" (*vị trí không gian*). Vào năm 2003, Mezaris và cộng sự đã giới thiệu hệ thống tra cứu ảnh dựa trên bản thể đối tượng. Trong hệ thống này, mỗi vùng của ảnh được mô tả bởi màu trung bình trong không gian màu *lab*, vị trí của nó trong trục dọc và trục ngang, kích cỡ và hình dạng của nó. Bản thể đối tượng được minh hoạ bởi hình 1-2.



Hình 1-2: Sơ đồ minh họa bản thể đối tượng

Vậy hệ thống đó đã hoạt động như thế nào?. Mấu chốt ở đây là việc lượng tử hoá đặc trưng kết cấu và màu sắc. Trong hệ thống máy tính hiện nay, đã có hàng triệu màu đã được định nghĩa. Nhưng với con người thì chỉ có khoảng từ 10 đến 20 màu là được đặt tên. Những mô hình tên màu biểu diễn mối liên quan giữa không gian màu số với tên màu được con người sử dụng.

Vào năm 1982, Berk và các cộng sự đã đề xuất một hệ thống tên màu nổi tiếng là "*CNS*" – (Color Naming System). Hệ thống này lượng tử hoá không gian màu HSL (Hue-Saturation-Luminance) thành 627 màu khác nhau. Ý tưởng chính ở đây là lượng tử giá trị "*Hue*" vào một tập các màu cơ bản. *Saturation* và *Luminance* thì được lượng tử hoá vào hai "*bin*" khác nhau mà để điều chỉnh độ tươi và độ sáng của màu. Một bộ tên màu đầu đủ trong "*CNS*" là: đỏ, cam, nâu, vàng, xanh lá cây, xanh dương và tím. Ngoài ra, với việc thêm vào các giá trị vô sắc như: đen, xám và trắng. Thì ta đã có được một mẫu 10 màu cơ bản.

Thêm nữa, để liên kết màu tới sự cảm nhận và trực quan cho việc tra cứu tranh, ảnh, thì một loại tương phản khác đã được định nghĩa như là: độ tương phản sáng-tối, độ tương phản ấm-lạnh, và độ tương phản bù, v.v…. Ví dụ là: màu vàng thì được gọi là "*ấm*", còn màu xanh lam thì được gọi là "*lạnh*". Còn một số nhà nghiên cứu khác thì cố gắng gán tên màu cho đối tượng liên quan trong cảnh thiên nhiên. Ví dụ: màu trắng thì gán với tuyết, mây, màu đỏ thì gán với mặt trời. Vì thế mà bằng cách này hệ thống đã làm giảm sự chênh lệch ngữ nghĩa và hỗ trợ truy vấn bằng từ khoá.

Tương tự như hệ thống đặt tên màu "*CNS*", chúng ta cũng cần xây dựng một hệ thống đặt tên kết cấu mà sẽ chuẩn hoá các mô tả và biểu diễn của kết cấu.

Tuy nhiên, cho đến nay vẫn chưa có một hệ thống đặt tên kết cấu nào có sẵn trong ứng dụng đời sống. Vì việc đặt tên cho kết cấu thực sự rất khó. Trong những bước đi đầu tiên để xây dựng một hệ thống đặt tên kết cấu, một số nhà nghiên cứu cố gắng xác định đặc trưng quan trọng mà con người dùng để nhận biết kết cấu. Dựa trên những thử nghiệm trực quan, vào năm 1993, Rao và các cộng sự đã chỉ ra rằng có ba thuộc tính quan trọng để con người nhận biết kết cấu. Đó là tính lặp lại, tính có hướng và độ phức tạp. Tuy nhiên, làm thế nào để thu được các đặc trưng này; và làm thế nào để có thể ánh xạ đặc trưng kết cấu mức thấp với ba yếu tố trên thì vẫn còn phải nghiên cứu thêm.

So với màu sắc thì kết cấu vẫn chưa được tìm hiểu sâu và mô hình hoá đầy đủ. Và có một hướng đi khác, thay vì việc sử dụng tên kết cấu như là từ khoá cho truy vấn. Thì một số nhà nghiên cứu lại cố gắng lượng tử hoá các cảm giác đặc trưng kết cấu thành các khoảng khác nhau và định nghĩa mô tả ý nghĩa kết cấu.

Cuối cùng, phương pháp bản thể đối tượng sẽ hoạt động tốt với một cơ sở dữ liệu ảnh đặc trưng với ngữ nghĩa đơn giản. Nhưng với một bộ các ảnh có nội dung khác nhau lớn hơn thì cần phải có những công cụ hiệu quả hơn để hệ thống có thể học được ngữ nghĩa.

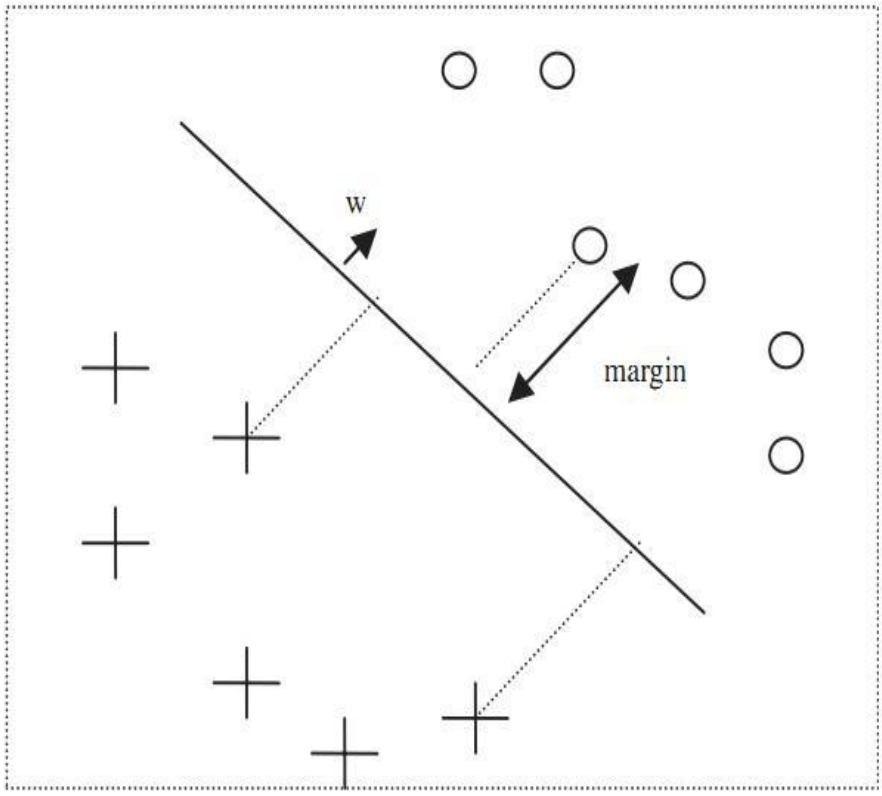
1.4.2 Kỹ thuật máy học

Trong một số trường hợp, để có thể suy ra đặc trưng ngữ nghĩa mức cao, thì hệ thống yêu cầu cần phải có một công cụ hình thức như là kỹ thuật máy học. Mà trong đó kỹ thuật máy học có hai loại là: học có giám sát và không có giám sát. Sau đây, chúng ta sẽ tìm hiểu kỹ thuật máy học có giám sát.

1.4.2.1 Học có giám sát

Học có giám sát dựa theo thuật toán *Support Vector Machines* (SVM) và phân lớp *Bayesian* thường được sử dụng để học các khái niệm mức cao từ các đặc trưng ảnh mức thấp. Với một nền tảng lý thuyết mạnh mẽ, *SVM* đã được sử dụng để nhận dạng đối tượng, phân loại text, v.v…., và được xem như là một giải pháp tốt cho việc học trong hệ thống tra cứu ảnh. Ban đầu, *SVM* được thiết kế cho việc phân lớp nhị phân. Giả sử rằng: có một tập dữ liệu huấn luyện $\{x_1, x_2, ..., x_n\}$, như là các *véc-tơ* trong không gian $X \subseteq R^d$ thuộc về hai lớp rời rạc với các nhãn của nó là $\{y_1, y_2, ……, y_n\}$ và $y_i \in \{-1, 1\}$. Sau đó, chúng ta có một mặt phẳng phân tách tối ưu (*Optimal separating plane*) nhằm cực đại hóa khoảng cách từ siêu phẳng đến các điểm dữ liệu gần nhất như hình 1-3.

Các véc-tơ nằm trên một mặt dưới sẽ được gắn nhãn là *-1*, còn các véc-tơ ở mặt trên sẽ được gắn nhãn là *+1*. "Hỗ trợ véc-tơ" đề cập đến các mẫu huấn luyện nằm gần với siêu phẳng nhất. Để học được nhiều khái niệm cho tra cứu ảnh, mỗi một bộ *SVM* sẽ được huấn luyện cho từng bộ khái niệm. Ví dụ cho việc dùng *SVM* trong chú thích ảnh. Trong giai đoạn huấn luyện, mỗi một mô hình *SVM* nhị phân sẽ được huấn luyện cho từng bộ khái niệm trong 23 bộ khái niệm được lựa chọn. Đến giai đoạn kiểm thử thì vùng dữ liệu chưa được gắn nhãn sẽ được đưa lần lượt vào từng bộ *SVM*. Cho đến khi mô hình *SVM* nào cho ra kết quả dương cao nhất thì mô hình đó sẽ thích hợp với vùng dữ liệu đó.



Hình 1-3: Sơ đồ mô tả kỹ thuật SVM

Một phương pháp học khác cũng được sử dụng rộng rãi là phân lớp *Bayesian*. Vào năm 2001, Ailaya và các cộng sự đã sử dụng lớp phân loại nhị phân *Bayesian* để ghi lại các khái niệm mức cao của cảnh tự nhiên từ đặc trưng mức thấp. Cơ sở dữ liệu ảnh sẽ tự động phân loại ảnh vào trong một loại chung là *nội cảnh/ngoại cảnh*, sau đó các ảnh *ngoại cảnh* lại tiếp tục được phân loại tiếp vào trong loại *thành phố/phong cảnh*, v.v…. Vì thế mà vào năm 2001, Luo và Savakis đã dùng mạng *Bayesian* để phân loại ảnh *nội cảnh/ngoại cảnh*.

Một kỹ thuật học khác để học các khái niệm là mạng *nơ-ron*. Để sử dụng kỹ thuật này thì các khái niệm phải được phân chia thành 11 loại là: gạch, mây, lông thú, cỏ, đá băng, đường, đá, cát, da, cây và nước. Sau đó, một số lượng lớn dữ liệu đã được huấn luyện (chính là các đặc trưng mức thấp của vùng đã được phân đoạn) được đưa vào mạng *nơ-ron* phân lớp để thiết lập liên kết giữa đặc trưng mức thấp và ngữ nghĩa mức cao (các nhãn phân loại). Một bất lợi của kỹ thuật này là nó yêu cầu một số lượng lớn các dữ liệu đã được huấn luyện và độ phức tạp tính toán cao.

Ba thuật toán ở trên tồn tại hai nhược điểm:

- Cần một số lượng lớn các mẫu huấn luyện được gắn nhãn, và các dữ liệu này thì dễ bị lỗi.
- Tập huấn luyện thì phải được cố định suốt trong quá trình học và giai đoạn ứng dụng.

Do đó, nếu mà miền ứng dụng thay đổi, thì các mẫu được gắn nhãn mới phải được cung cấp để đảm bảo tính hiệu quả phân lớp.

Bên cạnh các thuật toán đã được nêu trên, kỹ thuật cây quyết định (*decision tree*) cũng được dùng để biểu diễn đặc trưng ngữ nghĩa. Một số phương pháp đã xây lên một cấu trúc cây bằng việc phân chia đệ quy không gian thuộc tính *input* vào trong một tập không gian không chồng lấp. Một tập luật quyết định có thể được biểu diễn bởi đường dẫn từ gốc cây đi đến ngọn. Vào năm 2001, Sethi và Coman đã sử dụng phương pháp cây quyết định *CART* để biểu diễn luật quyết định ánh xạ phân bố màu toàn cục (*HSV biểu đồ không gian màu*) trong một ảnh để chú thích văn bản (4 từ khóa: *Sunset, Marine, Arid images and Nocturne*). Còn MacArthur và các cộng sự của ông thì dùng phương pháp cây quyết định *C4.5* được dựa trên một tập ảnh liên quan cho truy vấn. Sau đó phương pháp này được sử dụng như một mô hình để phân lớp cơ sở dữ liệu ảnh vào hai lớp: liên quan và không liên quan. Thuật toán này được sử dụng trong vòng lặp phản hồi liên quan (*RF*) để cung cấp các ảnh liên quan cho người dùng gắn nhãn ở vòng lặp tiếp theo.

So với các phương pháp học khác, cây quyết định là khái niệm đơn giản, hiệu quả với các đặc trưng đầu vào không đầy đủ và nhiễu. Thêm vào đó, cây quyết định có thể dễ dàng chuyển thành một tập quy tắc có thể tích hợp vào một hệ thống chuyên gia để đưa ra những quyết định tự động. Tuy nhiên, nhược điểm của phương pháp này là thiếu tính *mô-đun*, nếu mà sử dụng trong việc học khái niệm mức cao trong tra cứu ảnh và các bài toán cơ sở.

Các phương pháp sau này được phát triển bởi các tác giả ở trên đã phần nào khắc phục được những nhược điểm đó. Và thêm nữa, các phương pháp nêu trên sử dụng giá trị thuộc tính *input*, nhưng thông thường các đặc trưng ảnh mức thấp thì có giá trị liên tục. Mặc dù, một số thuật toán đã được thiết kế để rời rạc hóa các thuộc tính liên tục. Thế nhưng liệu có hay không các thuật toán được thiết kế ra để phân tách ý nghĩa của không gian đặc trưng ảnh thì cho đến nay vẫn chưa có lời giải.

1.4.2.2 Học không giám sát

Không giống như kỹ thuật học có giám sát là có sự hiện diện của kết quả trong quá trình học. Học không giám sát thì sẽ không cho kết quả đầu ra, mà nhiệm vụ chính được đặt ra là: làm thế nào để tổ chức hoặc phân cụm các đặc trưng đầu vào. Phân cụm ảnh là một kỹ thuật học không giám sát điển hình cho mục đích tra cứu. Kỹ thuật này cố gắng gom các dữ liệu ảnh giống nhau vào trong một cụm một cách tối đa, và giảm thiểu sự giống nhau giữa các cụm khác nhau. Mỗi cụm kết quả sẽ được liên kết với một nhãn lớp và ảnh trong một cụm thì sẽ tương tự nhau.

Phương pháp phân cụm *k-mean* truyền thống và các biến thể của nó thì thường được dùng để phân cụm ảnh. Vào năm 2001, Stan và Sethi đã sử dụng phương pháp phân cụm *k-mean* để áp dụng cho các đặc trưng ảnh mức thấp của một tập các ảnh huấn luyện. Sau đó, số liệu thống kê đo sự thay đổi của mỗi cụm được sử dụng để lấy một tập hợp các ánh xạ giữa các đặc trưng mức thấp và các đặc tính văn bản tối ưu (*như là từ khoá*) của mỗi cụm tương ứng. Các quy tắc ánh xạ có thể được sử dụng để thêm ảnh chưa được gắn chỉ số vào trong cơ sở dữ liệu.

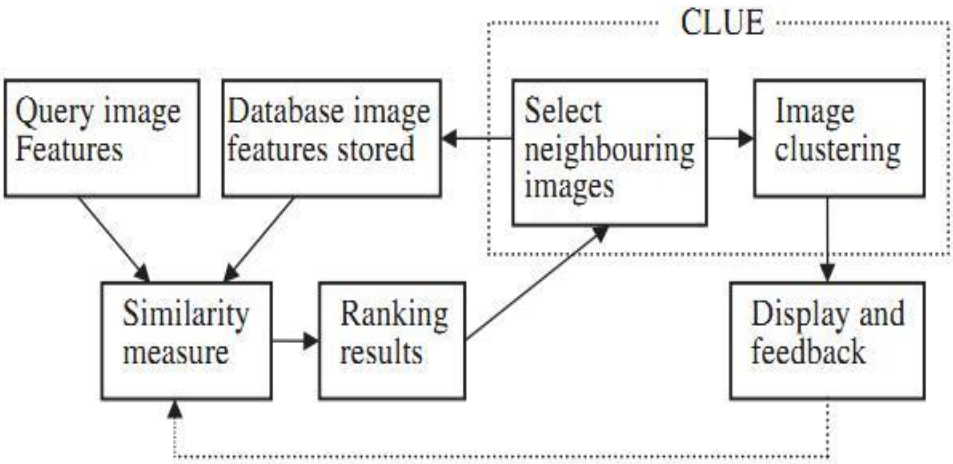
Vào năm 2004, Jin và các cộng sự đã sử dụng một phương pháp để chú thích ảnh trong cơ sở dữ liệu một cách tự động cho mục đích tra cứu. Đầu tiên, hệ thống sẽ phân cụm vùng ảnh vào trong một cụm mà sử dụng một biến thể của phương pháp phân cụm *k-mean*. Phương pháp đó được gọi là ràng buộc từng cặp *k-mean* (*PCK-mean*). Số cụm được thiết lập để thực nghiệm là 300. Sau đó, xác suất hậu nghiệm của mọi khái niệm (59 khái niệm được định nghĩa cho cơ sở dữ liệu ảnh) được đưa cho một vùng sử dụng phương pháp "*semi-naı¨veBayesian*". Phương pháp "*semi-naı¨veBayesian*" được Jin và các cộng sự sử dụng vào năm 2004. Do đó, một ảnh mới có thể được chú thích bằng việc chọn các khái niệm với xác suất cao nhất.

Do sự phân bố phức tạp của dữ liệu ảnh (các điểm dữ liệu được lấy mẫu từ đa dạng phi tuyến tính), mà các phương pháp truyền thống như phân cụm *k-mean* thường không thể phân chia tốt các ảnh với các khái niệm khác nhau. Để giải quyết vấn đề này, vào năm 2000, Shi và Malik đã đề xuất một phương pháp phân cụm

quang phổ *Normalized cut* (*NCut*). Phương pháp này đã được sử dụng thành công trong một vài ứng dụng như phân đoạn ảnh, phân cụm ảnh.

Vào năm 2003, Chen và các cộng sự đã trình bày một phương pháp tên là "*CLUE*" để giảm khoảng cách ngữ nghĩa trong CBIR. Không giống như các hệ thống CBIR khác mà chỉ hiển thị các ảnh trùng khớp trên cùng cho người dùng. Thì hệ thống này cố gắng tra cứu ngữ nghĩa một cách tự động và gắn kết các cụm ảnh. Cho một truy vấn ảnh, một tập các ảnh đích tương tự cho truy vấn được lựa chọn như là láng giềng của truy vấn. Dựa trên giả thuyết rằng, các ảnh có ngữ nghĩa giống nhau thì có xu hướng bị phân cụm. Phương pháp *Ncut* được sử dụng cho việc phân cụm các ảnh đích vào trong các lớp ngữ nghĩa khác nhau. Sau đó, hệ thống hiển thị cụm ảnh và điều chỉnh mô hình đo độ tương tự theo phản hồi của người dùng. Hình 1-4 là lược đồ cho hệ thống đó.

Phương pháp *Ncut* không thể tạo ra một hàm ánh xạ rõ ràng. Để giải quyết các điểm dữ liệu mới, thì độ tương tự giữa các điểm đó và tất cả dữ liệu huấn luyện phải được đo. Việc tính toán này có thể bị phức tạp do kích thước lớn của bộ huấn luyện.



Hình 1-4: Lược đồ mô tả phương pháp "CLUE"

Để giải quyết nhược điểm này thì vào tháng 10 năm 2004, Zheng và các cộng sự đã đề xuất ra một phương pháp phân cụm lưu trữ cục bộ (*locality preserving clustering-LPC*) cho việc phân cụm ảnh. Kết quả thí nghiệm cho thấy phương pháp *LPC* có thể cung cấp việc tra cứu chính xác tương đương với phương pháp *Ncut*, nhưng lại có hiệu năng tính toán cao hơn. Thêm nữa là, kết quả tra cứu của *LPC* được chứng minh là chính xác hơn phương pháp phân cụm *k-mean*.

1.4.2.3 Kỹ thuật tra cứu ảnh nhận dạng đối tượng

Nhận dạng đối tượng trong ảnh là một bài toán quan trọng trong thị giác máy tính với các ứng dụng trong chú thích ảnh, giám sát và tra cứu ảnh. Các thuật toán nhận dạng đối tượng giám sát hoặc không có giám sát đã được phát triển gần đây để có thể tra cứu ảnh dựa trên ngữ nghĩa. Vào năm 2003, Fergus và các cộng sự đã giới thiệu một phương pháp học bất biến co dãn không giám sát (*unsupervised scale-invariant learning*) để học và nhận dạng mô hình lớp đối tượng từ các cảnh được phân cụm nhưng chưa được phân đoạn và chưa gắn nhãn. Trong phương pháp này, các đối tượng được mô hình hoá như là từng phần của đồ thị và một biểu diễn xác xuất được sử dụng cho hầu hết các khía cạnh của đối tượng như là: hình dạng, bề ngoài, khớp, độ co dãn tương đối. Trong quá trình nhận dạng, mô hình này được sử dụng trong kiểu *Bayesian* cho phân loại ảnh. Mô hình tự nhiên linh hoạt đã được chứng minh bằng kết quả tốt trong một loạt các bộ dữ liệu bao gồm: các lớp hình học ràng buộc (ví dụ như: khuôn mặt, xe hơi) và các đối tượng mềm dẻo (ví dụ như: động vật).

Có một thực tế rằng, hầu hết người dùng muốn tra cứu ảnh dựa trên các đối tượng trong ảnh. Vào tháng 8 năm 2004, Li và các cộng sự đã phát triển một phiên bản bán giám sát mới của thuật toán *EM* cho việc học sự phân chia của lớp đối tượng. Các ảnh được biểu diễn là bộ *véc-tơ* đặc trưng của nhiều loại vùng trừu tượng. Mỗi vùng trừu tượng được mô hình hoá như là một pha trộn của sự phân chia *Gaussian* trên không gian đặc trưng. Các vùng được sử dụng trong việc nhận dạng có thể đến từ các quá trình phân đoạn khác nhau. Các vùng đó được sử dụng thì gọi là "*vùng trừu tượng*". Một mấu chốt của hướng tiếp cận này là không cần biết vị trí của đối tượng trong ảnh. Các thí nghiệm trên một tập 860 ảnh đã chứng minh tính hiệu quả của phương pháp này.

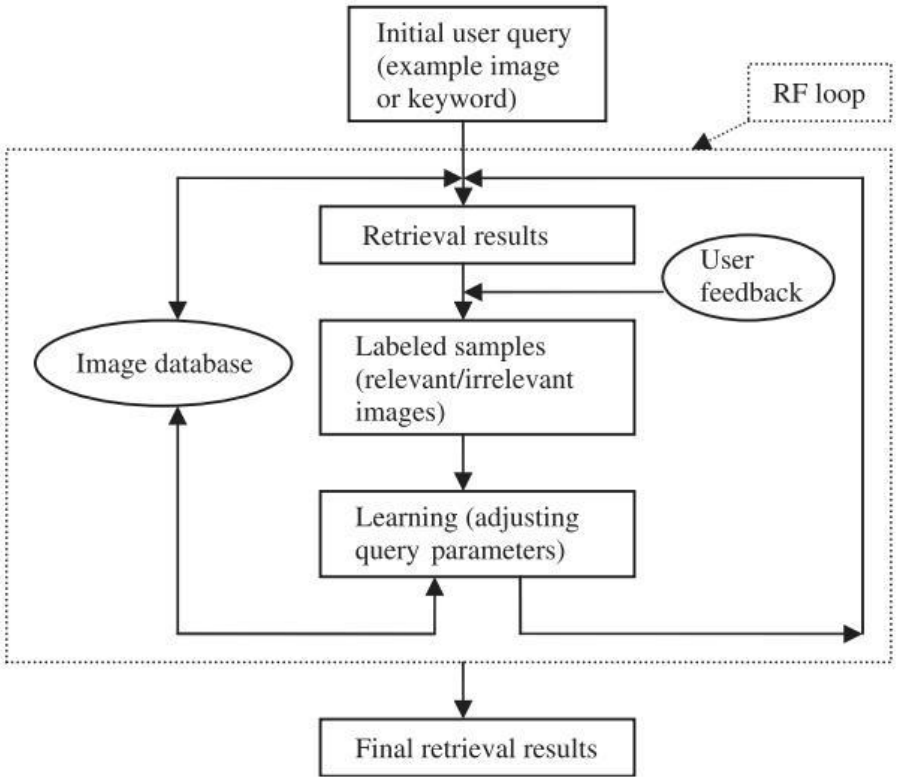
Vào năm 2005, Li và các cộng sự đã đề xuất phương pháp học lan truyền hoặc phân tách hai pha để học cách nhận biết đối tượng sử dụng nhiều loại đặc trưng. Mục tiêu của cách làm này là phát triển một phương pháp luận để phân loại ảnh ngoại cảnh. Pha lan truyền sẽ bình thường hoá mô tả độ dài ảnh, cái mà có thể tuỳ chỉnh số lượng của từng loại đặc trưng được trích chọn. Còn pha phân tách, một cách học phân loại mà ảnh sẽ được biểu diễn bởi mô tả độ dài cố định, và bao gồm đối tượng đích. Trong kết quả thực nghiệm của phương pháp này, bằng việc sử dụng màu sắc, kết cấu, và các cấu trúc đặc trưng đã cho thấy rằng hiệu suất tra cứu rất khả thi trên 31 loại phần tử đối tượng và 20 khái niệm mức cao.

1.4.3 Kỹ thuật phản hồi liên quan

Khái niệm phản hồi liên quan đã được giới thiệu trong tra cứu ảnh dựa trên nội dung từ khái niệm tra cứu thông tin dựa trên văn bản từ cuối những năm 90 và sau đó đã trở thành một công nghệ phổ biến cho *CBIR* để giảm khoảng cách ngữ nghĩa giữa đặc trưng mức thấp và các khái niệm ngữ nghĩa mức cao [3]. Nói chung, phản hồi liên quan nhằm mục đích cải thiện hiệu năng tra cứu bởi học với sự điều chỉnh của người dùng trên kết quả tra cứu. Trong cách này, hệ thống cần phải chạy thông qua một số vòng lặp. Trong mỗi vòng lặp, hệ thống trước tiên sẽ trả về một danh sách các ảnh kết quả đã được sắp xếp gần nhất với ảnh truy vấn dựa trên khoảng cách *Euclidean*. Sau đó, một số ảnh được đưa ra để người dùng gán nhãn liên quan hoặc không liên quan tới ảnh truy vấn. Sử dụng các ảnh đã được gán nhãn này như là các mẫu, các kỹ thuật học máy sẽ được sử dụng để học và phân lớp các ảnh trong cơ sở dữ liệu thành hai lớp liên quan và không liên quan. Bằng việc học một cách liên tục thông qua sự tương tác với người dùng cuối, phản hồi liên quan đã cho thấy việc tăng hiệu năng đáng kể trong hệ thống *CBIR* [4]. Một quá trình xử lý điển hình của *RF* trong *CBIR* được mô tả như sau:

1. Người dùng lựa chọn ảnh truy vấn. Sau đó, đặc trưng mức thấp của ảnh được trích chọn.
2. Hệ thống trả lại kết quả ảnh. Quá trình này có hai trường hợp:
   a. Pha ban đầu: Dựa vào đo độ tương tự của đặc trưng mức thấp giữa đặc trưng ảnh truy vấn và đặc trưng ảnh trong cơ sở dữ liệu để xếp hạng ảnh kết quả.
   b. Trong các vòng lặp *RF*: Sử dụng hàm phân lớp để xếp hạng ảnh kết quả.
3. Người sử dụng sẽ quyết định chọn những ảnh kết quả có mức độ giống với ý định của mình nhất. Đó là những ảnh liên quan (*mẫu dương*) hay không liên quan (*mẫu âm*) với ảnh truy vấn.
4. Thuật toán máy học sẽ được áp dụng để học phản hồi của người dùng sử dụng các mẫu được gắn nhãn thu được từ vòng lặp đầu tiên đến vòng lặp hiện tại. Sau đó, quay lại bước 2.

Bước (2), (3) và (4) sẽ được lặp lại nhiều lần cho đến khi người dùng hài lòng với kết quả tra cứu. Hình 1-5 sẽ cho thấy cách hoạt động của *RF* trong *CBIR*.



Hình 1-5: Sơ đồ mô tả phương pháp RF

Có nhiều cách tiếp cận trong bước (4), mà từ quan điểm máy học chung, về bản chất *RF* là một bài toán phân lớp nhị phân. Trong đó, các ảnh mẫu được cung cấp bởi người dùng được dùng để huấn luyện một lớp phân loại. Lớp này sẽ được sử dụng để phân loại ảnh trong cơ sở dữ liệu thành những loại liên quan đến truy vấn và phần còn lại thì không. Tuy nhiên, *RF* rất khác từ các bài toán phân lớp truyền thống. Bởi vì, những phản hồi được cung cấp từ người dùng thì lại bị giới hạn trong hệ thống tra cứu ảnh trên thực tế. Vì vậy, một phương pháp học mẫu nhỏ sẽ là hướng đi rất hứa hẹn trong *RF*.

1.4.4 Mẫu ngữ nghĩa

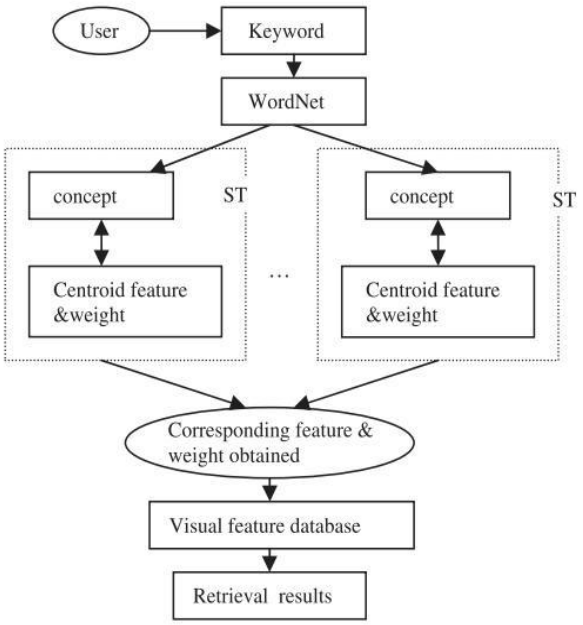
Mẫu ngữ nghĩa (*Semantic template*), mặc dù kỹ thuật này chưa được sử dụng rộng rãi như các kỹ thuật đã đề cập ở trên, nhưng lại là một cách tiếp cận đầy hứa hẹn trong việc tra cứu ảnh dựa trên ngữ nghĩa. Mẫu ngữ nghĩa là một ánh xạ giữa các khái niệm mức cao và các đặc trưng thị giác mức thấp. Mẫu ngữ nghĩa được định nghĩa như là khái niệm đặc trưng "*biểu diễn lại*" được tính toán từ một bộ sưu tập các ảnh mẫu. Trong một số hệ thống, biểu tượng hay các ảnh mẫu cũng được cung cấp cho sự tiện dụng truy vấn của người dùng.

Năm 1998, Chang và các cộng sự đã giới thiệu về ý tưởng mẫu ngữ cảnh (*semantic visual template*) để liên kết các đặc trưng ảnh mức thấp tới các khái niệm

mức cao trong tra cứu video. Một mẫu ảnh là một bộ các biểu tượng hoặc ví dụ về *cảnh/đối tượng* để biểu diễn các khái niệm theo quan điểm cá nhân như là : cuộc họp, hoàng hôn. Các đặc trưng véc-tơ của ví dụ về *cảnh/đối tượng* thì được trích chọn cho quá trình truy vấn. Để sinh các mẫu ngữ nghĩa, đầu tiên người dùng sẽ định nghĩa mẫu cho các khái niệm đặc biệt bằng cách xác định các đối tượng và rằng buộc không gian và thời gian. Trọng số thì được gán cho từng đặc trưng của từng đối tượng. Quá trình truy vấn ban đầu này được cung cấp cho hệ thống. Thông qua sự tương tác với người dùng, hệ thống cuối cùng sẽ hội tụ với một nhóm nhỏ truy vấn điển hình mà có sự trùng khớp "*tốt nhất*" (*độ chính xác cao*) các khái niệm trong tâm trí người dùng.

Thế hệ của phương pháp *SVT* của Chang và các cộng sự phụ thuộc vào sự tương tác với người dùng và yêu cầu người dùng phải có sự hiểu biết chuyên sâu về các đặc trưng ảnh. Điều này gây trở ngại cho các ứng dụng mà người dùng bình thường hay sử dụng. So với việc này, thì vào năm 1999, Zhuang và các cộng sự đã sử dụng một hệ thống khác để sinh mẫu ngữ nghĩa một cách tự động trong quá trình xử lý phản hồi liên quan, dựa trên những hiểu biết rằng phản hồi liên quan là một quá trình mà người dùng cụ thể hóa truy vấn ngữ nghĩa. Thứ nhất, người dùng gửi ảnh truy vấn với một khái niệm (*từ khóa*) đại diện cho ảnh. Sau đó lặp lại nhiều lần, hệ thống sẽ trả về một số hình ảnh liên quan cho người dùng. Đặc trưng trọng tâm của những ảnh đó sẽ được tính toán và sử dụng như đại diện của các khái niệm truy vấn. Sau đó, *ST* sẽ định nghĩa là *ST={C, F, W}* với khái niệm truy vấn *C*, đặc trưng trọng tâm *F* thu được, và trọng số *W* được áp dụng cho đặc trưng véc-tơ.

Năm 1990, Miller và các cộng sự đã giới thiệu hệ thống *WorldNet* để xây dựng một mạng lưới các mẫu ngữ nghĩa. Trong quá trình tra cứu, một khi người dùng gửi khái niệm truy vấn (*từ khóa*), thì hệ thống có thể tìm một mẫu ngữ nghĩa tương ứng, và sử dụng *F* và *W* tương ứng để tìm những ảnh tương tự. Quá trình tra cứu được thể hiện ở hình 1-6. Người dùng không thể thấy được việc sinh mẫu, và sử dụng hệ thống mà không cần bất kỳ kiến thức nào về đại diện đặc trưng.



Hình 1-6: Hệ thống tra cứu ảnh Worldnet sử dụng Semantic template

Một công việc liên quan khác được giới thiệu bởi Smith và Li vào năm 1998. Họ sử dụng một công nghệ được gọi là *CRT* để giải mã ngữ nghĩa ảnh. Công nghệ *CRT* được định nghĩa là vùng sắp xếp không gian nguyên mẫu trong ảnh. Cho một lớp ngữ nghĩa, một tập các ảnh mẫu được thu thập. Đầu tiên, hệ thống sẽ phân đoạn mỗi ảnh thành các vùng màu đồng nhất và trích chọn năm chuỗi bằng cách quét ảnh theo chiều dọc. Sau đó, hệ thống hợp nhất các vùng chuỗi bằng cách đếm tần số *CRT* trong tập hợp các khu vực chuỗi thu được từ tất cả các ảnh mẫu. Bằng việc kết hợp *CRT* từ mỗi lớp ngữ nghĩa tạo thành một thư viện *CRT*. Mô tả ngữ nghĩa của những ảnh chưa biết có thể được tạo ra bằng cách đối sánh sự sắp xếp của các khu vực ảnh cho thư viện *CRT*. Các thí nghiệm với một tập 10 lớp ngữ nghĩa (*bãi biển, tòa nhà, cua, thợ lặn, v.v...*) đã chứng minh rằng phương pháp này cải thiện độ chính xác tra cứu so với các phương pháp truyền thống sử dụng biểu đồ màu và đặc trưng kết cấu.

1.4.5 Tra cứu ảnh web

Chúng ta phân loại tra cứu ảnh web như là một trong các công nghệ mới trong tra cứu ảnh mức cao, hơn là một miền ứng dụng đặc biệt. Vì nó có một số khác biệt kỹ thuật từ tra cứu ảnh trong các ứng dụng khác. Một ưu điểm trong tra cứu ảnh web là một số thông tin bổ sung trên web có sẵn để tạo điều kiện tra cứu ảnh dựa trên ngữ nghĩa. Ví dụ, một file ảnh chứa đường dẫn thường có một cấu trúc phân cấp rõ ràng bao gồm một số thông tin về ảnh như loại ảnh. Ngoài ra, các tài liệu *HTML* cũng chứa một số thông tin hữu ích trong tiêu đề ảnh, *ALT-tag, các*

văn bản mô tả xung quanh ảnh, đường siêu liên kết, v.v… Tuy nhiên, thông tin đó chỉ có thể chú thích ảnh thành một số mở rộng.

Các công cụ tìm kiếm ảnh web có sẵn như là *Google* và *AltaVista* thì tìm kiếm ảnh chỉ dựa trên dấu hiệu từ khóa. Dù cho phương pháp này có thể tìm nhiều ảnh liên quan, thế nhưng độ tra cứu chính xác các ảnh lại rất kém. Vì chúng không thể xác nhận rằng có thực sự các ảnh được tra cứu có chứa các khái niệm truy vấn hay không?. Và kết quả là người dùng phải duyệt qua toàn bộ danh sách để tìm những ảnh mong muốn. Đây là một quá trình tốn nhiều thời gian vì các kết quả được trả về luôn luôn chứa nhiều chủ đề bị pha trộn với nhau. Để cải thiện hiệu năng tra cứu ảnh web, các nhà nghiên cứu đang nỗ lực để kết hợp các thông tin văn bản với nội dung ảnh trực quan.

Vào năm 2004, Feng và các cộng sự đã giới thiệu một phương pháp tên là *bootstrapping* đồng huấn luyện *framework*. Phương pháp này được sử dụng để chú thích ảnh web một cách tự động với một bộ các khái niệm tra cứu. Hệ thống khai thác các dấu hiệu từ cả hai loại là: văn bản *HTML* và đặc trưng ảnh trực quan. Và sau đó, phát triển hai phân loại độc lập dựa trên văn bản và đặc trưng ảnh trực quan tương ứng. Các kết quả thử nghiệm đã sử dụng một tập 15 các khái niệm tiền định nghĩa để biểu diễn hiệu suất thực của hệ thống. Tuy nhiên, do sự thiếu chính xác trong trích chọn thông tin văn bản, hiệu suất của một số khái niệm không được thoả mãn.

Vào năm 2004, Cai và các cộng sự đã phát triển một hệ thống tra cứu ảnh web đầy hứa hẹn tên là *MSRA* (*Microsoft Research Asia*). Mục đích của hệ thống này là phân cụm kết quả tìm kiếm của *engine* tìm kiếm ảnh web thông thường, sao cho người dùng có thể tìm thấy hình ảnh mong muốn một cách nhanh chóng. Trước tiên, một thuật toán phân đoạn dựa trên trực quan thông minh được thiết kế để phân đoạn một trang web thành các khối. Từ các khối có chứa ảnh này, văn bản và thông tin đường *link* của ảnh có thể được trích chọn một cách chính xác. Sau đó, một đồ thị ảnh được xây dựng bằng cách sử dụng các kỹ thuật phân tích liên kết mức-khối. Do đó, trong mỗi ảnh, ta thu được có ba loại: đại diện dựa trên đặc trưng trực quan, đại diện dựa trên đặc trưng kết cấu và đại diện dựa trên đồ thị.

Kết quả thử nghiệm ban đầu đã cho thấy rằng bằng cách kết hợp kết cấu và đại diện dựa trên đồ thị cho cụm ảnh, hệ thống có thể tra cứu cấu trúc ngữ nghĩa của ảnh web. Kết quả tìm kiếm được nhóm thành cụm vào trong các loại ngữ nghĩa khác nhau. Đối với mỗi loại, một số ảnh được chọn làm ảnh đại diện, để người dùng có thể hiểu chủ đề chính của các kết quả tìm kiếm một cách nhanh chóng.

Những ảnh trong mỗi loại này sau đó được tổ chức lại dựa trên các đặc trưng thị giác của chúng để làm cho cụm trực quan hơn trong mắt người dùng. Tuy nhiên, một đánh giá thử nghiệm toàn diện cần phải được thực hiện để kiểm tra tính hiệu quả của kỹ thuật này.

1.5 Các lĩnh vực ứng dụng của tra cứu ảnh dựa trên nội dung

Ứng dụng của tra cứu ảnh dựa trên nội dung có rất nhiều trong đời sống xã hội, phục vụ cho nhiều mục đích khác nhau, nhằm xác nhận, tra cứu thông tin. Nhờ đó mà giảm bớt công việc của con người, nâng cao hiệu suất làm việc, ví dụ như: Album ảnh số của người dùng, ảnh y khoa, bảo tàng ảnh, tìm kiếm nhãn hiệu, mô tả nội dung video, truy tìm ảnh tội phạm, hệ thống tự nhận biết điều khiển luồng giao thông… Một vài hệ thống lớn đại diện cho các lĩnh vực bao gồm :

- Hệ thống truy vấn ảnh theo nội dung (*Query By Image Content*) được nghiên cứu và phát triển bởi nhóm nghiên cứu *Visual Media Management* thuộc công ty IBM, đây là một hệ thống tra cứu ảnh thương mại được phát triển từ rất sớm. Hiện nay, hệ thống này hỗ trợ một vài đo độ tương tự cho ảnh như: trung bình màu sắc, lược đồ màu sắc và kết cấu. Công nghệ sử dụng trong hệ thống bao gồm 2 phần chính là: đánh chỉ số và tìm kiếm. Hơn nữa, hệ thống này còn cung cấp vài cách tiếp cận truy vấn theo đơn đặc trưng, đa đặc trưng và đa giai đoạn.
- Hệ thống *Visual SEEK* tại trường đại học Columbia. Hệ thống cho phép người dùng nhập vào truy vấn, sử dụng các đặc trưng mức thấp của hình ảnh như: màu sắc, bố cục không gian và kết cấu. Các đặc trưng đó được mô tả theo màu sắc và biến đổi *Wavelet* dựa trên đặc trưng kết cấu.
- Hệ thống *NeTra* sử dụng các đặc trưng của ảnh: Màu sắc, hình dạng, kết cấu, vị trí không gian.
- Ngoài ra, còn một số hệ thống khác như: Virage system, Stanford SIMPLICity system, NEC PicHunter system, v.v…

CHƯƠNG 2: TRA CỨU ẢNH DỰA TRÊN NỘI DUNG VỚI PHẢN HỒI LIÊN QUAN

2.1 Giới thiệu phương pháp phản hồi liên quan

So sánh với các công nghệ dựa trên đặc trưng mức vùng, thì công nghệ dựa trên đặc trưng mức đối tượng tập trung vào chi tiết nội dung thông tin hơn. Loại công nghệ này trước tiên áp dụng phương pháp phân đoạn ảnh để thu được các đối tượng độc lập trong một ảnh. Công nghệ này trích xuất các đối tượng đặc trưng trực quan là: màu sắc, kết cấu, hình dạng, v.v… để tạo thành véc-tơ đặc trưng trực quan mức thấp cho một ảnh. Cuối cùng, công nghệ này áp dụng thuật toán đối sánh trên những đặc trưng mức đối tượng để tính toán số điểm liên quan cuối cho từng ảnh trong cơ sở dữ liệu. Vào năm 2001, Wang và các cộng sự đã đề xuất phương pháp *SIMPLIcity* để áp dụng cho các đặc trưng mức đối tượng như: màu sắc, kết cấu, hình dạng trong hệ thống *CBIR*, và đã chứng minh tính hiệu quả của công nghệ này. Tuy nhiên, việc phân đoạn ảnh vẫn là chủ đề nghiên cứu đầy thách thức trong lĩnh vực thị giác máy tính. Đó không phải là một giải pháp phân đoạn phổ thông cho tất cả các loại ảnh. Do đó, hệ thống dựa trên đặc trưng mức đối tượng sẽ phải chịu việc giảm chất lượng ảnh đã được phân đoạn.

Kỹ thuật dựa trên phản hồi liên quan (*RF*) được giới thiệu vào năm 2007 bởi Liu và các cộng sự. Đây là kỹ thuật học trực tuyến có giám sát mà được sử dụng rộng rãi trong hệ thống *CBIR* để khắc phục khoảng cách ngữ nghĩa. *RF* sẽ thay đổi nhiều lần thông tin mô tả truy vấn (*đặc trưng, mô hình đối sánh, metrics hoặc bất kỳ tri thức meta*) như là hồi đáp phản hồi của người dùng trên kết quả tra cứu. Vì thế, kỹ thuật này sẽ học các truy vấn gần với nó nhất và trả về nhiều ảnh mà người dùng mong muốn (cải thiện độ chính xác tra cứu) sau mỗi vòng. Hệ thống *CBIR* tương tác dựa trên *RF* lần đầu tiên được đề xuất vào năm 1998 bởi Rui và các cộng sự. Người dùng cung cấp sự lựa chọn trên các ảnh đã tra cứu trong lần lặp tra cứu trước. Kỹ thuật này được sử dụng để vượt qua hai nhược điểm chính trong hệ thống không dựa trên *RF*.

- Khoảng cách ngữ nghĩa giữa khái niệm ngữ nghĩa mức cao và đặc trưng trực quan ảnh mức thấp.
- Nhận thức chủ quan của con người về nội dung trực quan.

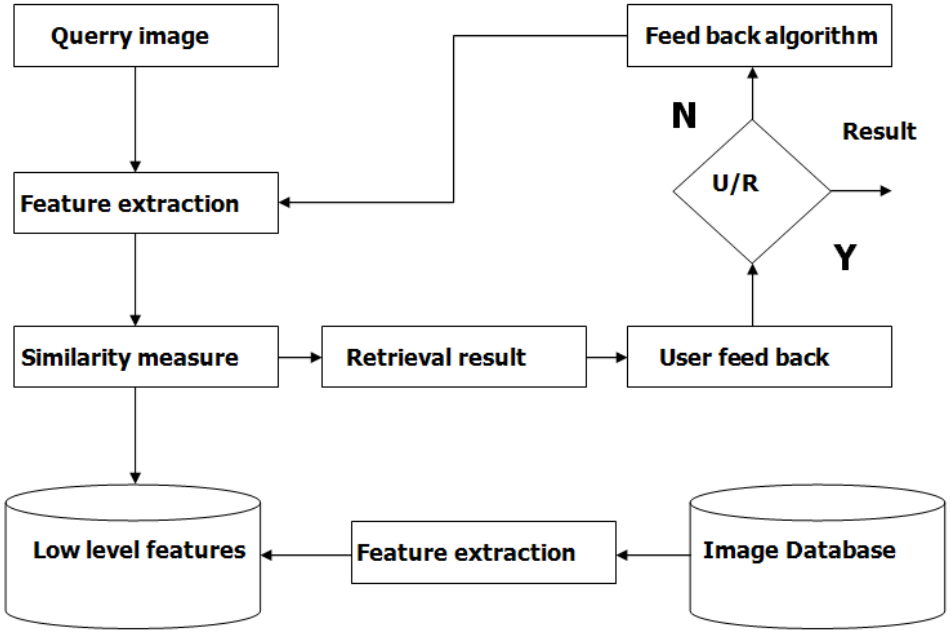
Đặc biệt, hệ thống dựa trên *RF* lần đầu tiên cập nhật trọng số đặc trưng tương ứng một cách tự động để nắm bắt mục đích của người dùng trong truy vấn và nhận

thức chủ quan sau mỗi vòng lặp truy vấn. Kết quả là, hệ thống *CBIR* dựa trên *RF* đã cải thiện hiệu năng tra cứu so với các hệ thống không dựa trên *RF* khác.

Người dùng đóng một vai trò quan trọng trong hệ thống *CBIR* dựa trên *RF*. Những phản hồi chính xác từ người dùng sẽ làm tăng hiệu năng của hệ thống *CBIR* để nắm bắt ý định tra cứu của họ. Kết quả là, các nhà nghiên cứu đang tập trung áp dụng thuật toán học trên *RF* để cải thiện hiệu năng tra cứu. Các thuật toán học đó có thể được phân loại vào trong hai kỹ thuật là: học ngắn hạn và học dài hạn. Việc lựa chọn chính xác thuật toán học nào thì lại phụ thuộc vào các ứng dụng tra cứu trong thực tế. Và chúng ta không thể trả lời chắc chắn rằng học ngắn hạn sẽ tốt hơn học dài hạn hoặc ngược lại.

2.2 Kiến trúc tổng quan của hệ thống

Hình vẽ 2-1 dưới đây mô tả sơ đồ tổng quát của tra cứu ảnh từ cơ sở dữ liệu sử dụng phản hồi liên quan. Ý tưởng cơ bản của phương pháp phản hồi liên quan là quay vòng truy vấn từ người dùng đến hệ thống cho đến khi nào ra được kết quả tối ưu nhất. Khi người dùng chọn **N** thì tức là người dùng đã cung cấp thêm thông tin cho hệ thống. Và hệ thống có thể sử dụng thuật toán phản hồi để cho ra kết quả tối ưu hơn nhờ vào lượng thông tin mà người dùng vừa cung cấp. Việc trích chọn ảnh từ cơ sở dữ liệu, hệ thống trên có thể trích chọn những đặc trưng véc-tơ của ảnh (những đặc trưng đó là: hình dạng, màu sắc, kết cấu, v.v…), và sau đó các đặc trưng véc tơ này sẽ được lưu lại trong cơ sở dữ liệu để được sử dụng cho việc truy vấn ảnh sau này. Khi hệ thống nhận được lệnh truy vấn từ phía người dùng, các đặc trưng véc tơ sẽ được trích chọn từ ảnh truy vấn, và được đối sánh với các đặc trưng đã được lưu trong cơ sở dữ liệu. Nếu như khoảng cách đặc trưng véc tơ của hai ảnh là đủ nhỏ, thì hệ thống sẽ trả kết quả hình ảnh từ cơ sở dữ liệu ra phía người dùng.



Hình 2-1: Kiến trúc hệ thống tra cứu ảnh dựa trên nội dung với phản hồi liên quan

Việc tìm kiếm ảnh thường dựa trên sự tương tự hơn là trích chọn chính xác, và kết quả tra cứu sẽ được đưa ra cho người dùng. Sau đó, người dùng đưa ra các thông tin phản hồi trong một bản mẫu "*Các quyết định liên quan*" thể hiện thông qua kết quả tra cứu. "*Quyết định liên quan*" đánh giá kết quả dựa trên ba giá trị. Ba giá trị đó là: liên quan, không liên quan, và không quan tâm. "*Liên quan*" nghĩa là ảnh có liên quan đến truy vấn của người dùng. "*Không liên quan*" có nghĩa là ảnh không có liên quan đến truy vấn người dùng. Còn "*không quan tâm*" nghĩa là người dùng không cho biết bất kỳ điều gì về ảnh. Nếu phản hồi của người dùng là có liên quan, thì vòng lặp phản hồi sẽ tiếp tục hoạt động cho đến khi người dùng hài lòng với kết quả tra cứu.

Như hình 2-1 mô tả cấu trúc của hệ thống phản hồi liên quan. Trong hệ thống đó có các khối chính là: cơ sở dữ liệu ảnh, trích chọn đặc trưng, đo độ tương tự, phản hồi từ người dùng, và thuật toán phản hồi.

2.2.1 Trích chọn đặc trưng

Trích chọn đặc trưng liên quan đến việc trích chọn các thông tin có ý nghĩa từ ảnh. Vì vậy, nó làm giảm việc lưu trữ cần thiết, và do đó hệ thống sẽ trở nên nhanh hơn và hiệu quả trong CBIR. Khi đặc trưng được trích chọn, chúng sẽ được lưu trữ trong sơ sở dữ liệu để sử dụng trong lần truy vấn sau này. Mức độ mà một máy tính có thể trích chọn thông tin có ích từ ảnh là vấn đề then chốt nhất cho sự tiến bộ của hệ thống diễn giải hình ảnh thông minh. Một trong những ưu điểm lớn nhất của

trích chọn đặc trưng là: nó làm giảm đáng kể các thông tin (*so với ảnh gốc*) để biểu diễn một ảnh cho việc hiểu nội dung của ảnh đó. Kỹ thuật này đã đóng góp rất lớn cho các hướng tiếp cận khác nhau để phát hiện các loại đặc trưng trong ảnh. Những đặc trưng này có thể được phân loại như là đặc trưng toàn cục và đặc trưng cục bộ. Các đặc trưng phổ biến nhất mà được sử dụng là màu sắc, kết cấu và hình dạng.

- Đặc trưng toàn cục: Đặc trưng toàn cục phải được tính toán trên toàn bộ ảnh. Ví dụ, mức độ màu xám trung bình, biểu đồ về cường độ hình dạng, v.v… Ưu điểm của việc trích chọn toàn cục là nó cho ra cả trích chọn đặc trưng và tính toán độ tương tự một cách nhanh chóng. Tuy nhiên, chúng có thể bị tràn qua cho vị trí và do đó thì thất bại cho việc nhận dạng đặc trưng trực quan quan trọng. Để gia tăng cái thô cho biến đổi không gian, chúng ta có thể tìm hiểu trích chọn đặc trưng cục bộ.
- Đặc trưng cục bộ: Trong đặc trưng toàn cục, việc trích chọn đặc trưng được tính toán trên toàn bộ ảnh. Tuy nhiên, đặc trưng toàn cục không thể xử lý tất cả các phần của ảnh có đặc điểm khác nhau. Do vậy, chúng ta cần trích chọn các đặc trưng cục bộ của ảnh. Các đặc trưng đó có thể được tính toán trên các kết quả của phân đoạn ảnh và thuật toán phát hiện đường biên. Vì thế, tất cả chúng đều dựa trên một phần của ảnh với một số tính chất đặc biệt.
- Điểm quan tâm: Trong việc tính toán đặc trưng cục bộ, việc trích chọn đặc trưng ảnh được giới hạn trong một tập con các điểm ảnh. Các điểm quan tâm, tập các điểm quan tâm được gọi là điểm nổi bật. Điểm nổi bật là những điểm có biến thiên cao trong đặc trưng của vùng lân cận điểm ảnh cục bộ. Nhiều hệ thống *CBIR* trích chọn những điểm nổi bật. Năm 2004, Rouhollah và các cộng sự đã định nghĩa điểm nổi bật có mặt trong tra cứu ảnh dựa trên nội dung như là một nhiệm vụ của *CBIR*, nơi mà người dùng chỉ quan tâm đến một phần của ảnh, và phần còn lại là không liên quan. Ví dụ, chúng ta có thể tham khảo một số đặc trưng cục bộ như là ảnh nguyên bản, đường tròn, đường nét, *texel* (các phần tử tập trung ở một khu vực kết cấu), hoặc các đặc trưng cục bộ khác, hình dạng của đường nét, v.v…

2.2.2 Đo độ tương tự

Trong độ đo tương tự, đặc trưng véc-tơ ảnh trong truy vấn và đặc trưng véc-tơ ảnh trong cơ sở dữ liệu được đối sánh bằng cách sử dụng khoảng cách *metric*. Các hình ảnh được xếp hạng dựa trên giá trị khoảng cách. Vào năm 2003, Manesh và các cộng sự đã đề xuất phương pháp đo độ tương tự cho việc đối sánh chi tiết 9

metric khác nhau như: Manhattan, weighted mean-variance, Euclidean, Chebychev, Mahanobis, v.v… khoảng cách cho tra cứu kết cấu ảnh với đánh giá thực nghiệm. Họ nhận thấy rằng số liệu khoảng cách Canberra and Bray-Curtis thực hiện tốt hơn các số liệu khoảng cách khác.

2.2.3 Phản hồi từ người dùng

Sau khi có kết quả tra cứu, người dùng cung cấp phản hồi về các kết quả liên quan hoặc không liên quan. Nếu kết quả là không liên quan thì vòng lặp phản hồi sẽ được lặp lại nhiều lần cho đến khi người dùng hài lòng.

2.3 Các phương pháp phản hồi liên quan

Trong các hệ thống *CBIR* với phản hồi liên quan, người dùng đóng một vai trò quan trọng. Các thông tin phản hồi chính xác từ người dùng sẽ góp phần làm tăng đáng kể hiệu năng của hệ thống tra cứu. Do vậy, các nhà nghiên cứu đã tập trung áp dụng các kỹ thuật học máy trên những phản hồi của người dùng để cải thiện hiệu năng tra cứu. Kỹ thuật cập nhật truy vấn và kỹ thuật học thống kê là những kỹ thuật được sử dụng phổ biến trong các hệ thống *CBIR* với phản hồi liên quan.

2.3.1 Kỹ thuật cập nhật truy vấn

Kỹ thuật cập nhật truy vấn cải thiện việc biểu diễn chính truy vấn bằng cách sử dụng thông tin được gắn nhãn chủ quan của người dùng. Các ví dụ của kỹ thuật cập nhật truy vấn bao gồm cập nhật trọng số truy vấn di chuyển truy vấn, và mở rộng truy vấn.

Cập nhật trọng số truy vấn làm thay đổi trọng số tương đối của các đặc trưng khác nhau trong biểu diễn truy vấn. Kỹ thuật cập nhật véc-tơ trọng số cho phép hệ thống học sự giải thích của người dùng về hàm khoảng cách. Ý tưởng trung tâm đằng sau phương pháp cập nhật trọng số rất là đơn giản và trực quan. Mỗi ảnh được đại diện bởi một véc-tơ đặc trưng N chiều. Nó có thể được xem như là một điểm trong không gian N chiều. Các chiều đặc trưng quan trọng để giúp tra cứu các ảnh liên quan sẽ được nâng cấp tầm quan trọng trong khi các chiều khác mà cản trở tiến trình này sẽ bị giảm tầm quan trọng. Vào năm 2004, Kushki và các cộng sự đã sử dụng kỹ thuật cập nhật trọng số để học ánh xạ tối ưu giữa đặc trưng trực quan mức thấp và khái niệm ngữ nghĩa mức cao của ảnh. Kỹ thuật này hoạt động bằng cách tinh chỉnh các trọng số (hoặc sự quan trọng) của từng thành phần đặc trưng hoặc bằng cách thay đổi đo độ tương tự một cách tương ứng. Cũng trong năm 2004, Muneesawang và cộng sự đã áp dụng kỹ thuật di chuyển truy vấn để cho phép người dùng thay đổi trực tiếp đặc trưng của ảnh truy vấn bằng cách chỉ định các

thuộc tính của các ảnh liên quan hoặc không liên quan được đánh dấu bởi người dùng. Có nghĩa là, các đặc trưng của nội dung ảnh truy vấn được thay đổi theo hướng biểu diễn ngữ nghĩa chính xác hơn được cung cấp bởi người dùng trong suốt quá trình tra cứu. Vào năm 2005, Widyantoro và các cộng sự đã áp dụng kỹ thuật mở rộng truy vấn để thêm vào một tập các ảnh liên quan mà không được gắn nhãn bởi người dùng để bù đắp cho sự thiếu hụt những ảnh đã được gắn nhãn bởi người dùng giúp hệ thống nắm bắt ý nghĩa của ảnh truy vấn một cách chính xác hơn.

2.3.2 Những kỹ thuật học thống kê

Kỹ thuật học thống kê đã cải thiện giới hạn phân loại giữa những ảnh liên quan và không liên quan hoặc dự đoán những ảnh liên quan mà chưa được gắn nhãn trong suốt quá trình huấn luyện. Các ví dụ của kỹ thuật học thống kê bao gồm học quy nạp và học chuyển đổi.

Học quy nạp được định nghĩa như là một quá trình tiếp thu tri thức bằng cách vẽ ra các suy luận quy nạp từ giáo viên hoặc môi trường cung cấp sự kiện. Đây là một quá trình liên quan đến hoạt động khái quát, biến đổi, hiệu chỉnh, tinh chỉnh biểu diễn tri thức. Phương pháp học quy nạp được áp dụng trong hệ thống *CBIR* nhằm tạo ra các bộ phân lớp khác nhau để phân tách thành ảnh có liên quan (*mẫu dương*) và không có liên quan (*mẫu âm*), và khái quát tốt hơn những ảnh chưa gán nhãn. Ở đây, những ảnh có liên quan và không có liên quan là nhãn ảnh tra cứu dương và âm một cách tương ứng bởi người dùng trong suốt phiên tra cứu. Các kỹ thuật học quy nạp điển hình bao gồm *Mạng neural, Học cây quyết định, học Bayesian, Boosting, support vector machine* (SVM), *học SVM mờ* (FSVM). Vào năm 2000, MacArthur và cộng sự đã sử dụng cây quyết định trong ứng dụng *CBIR*. Các ảnh liên quan và không liên quan được đánh dấu bởi người dùng được sử dụng để phân chia không gian đặc trưng cho đến khi tất cả các ví dụ trong một phân vùng là cùng lớp. Năm 2003, Su và các cộng sự đã cung cấp phản hồi liên quan và không liên quan từ người dùng vào bộ *Phân loại Bayesian*. Những ảnh liên quan được sử dụng để ước lượng một phân bố *Gaussian*. Phân bố này dùng để biểu diễn những ảnh mà người dùng mong muốn trong khi những ảnh không liên quan thì lại được sử dụng để duyệt lại việc xếp hạng những ứng cử đã được tra cứu. Năm 2001, Tong và cộng sự đã đề xuất một hệ thống *CBIR* với sự trợ giúp của *SVM* để học đường bao thích hợp sử dụng mẫu liên quan và không liên quan đã thu thập được từ vòng lặp tra cứu trước đó. Đường bao này sau đó được sử dụng để phân tách ảnh trong cơ sở dữ liệu thành hai phân vùng liên quan và không liên quan. Năm 2006, Wu và các cộng sự đã áp dụng *FSVM* để học đường bao quyết định để

phân tách ảnh huấn luyện dương và âm dựa trên các trọng số mờ tương ứng. Đường bao quyết định sau đó được dùng để phân chia cơ sở dữ liệu ảnh thành ảnh liên quan và không liên quan. Những ảnh liên quan với khoảng cách lớn nhất tới đường bao quyết định được coi như là những ảnh tương tự nhất với ảnh truy vấn. Năm 2004, Tieu và cộng sự đã đề xuất một hệ thống *CBIR* mà sử dụng kỹ thuật học "*boosting*" để sinh ra một số lượng lớn các đặc trưng chọn lọc cao cho việc nắm bắt nhiều dạng của khái niệm trực quan ảnh. Một loạt các phương pháp học yếu dựa trên một số lượng nhỏ các đặc trưng đã được huấn luyện trong suốt thời gian truy vấn. Bằng việc kết hợp các phân loại yếu, hệ thống cuối cùng thu được một bộ phân loại mạnh có độ tương quan tốt hơn với phân lớp lý tưởng.

Kỹ thuật học truyền dẫn khai thác mối liên quan của tất cả các ảnh cơ sở dữ liệu trong không gian đặc trưng và lan truyền độ xếp hạng của ảnh được gắn nhãn với ảnh chưa gắn nhãn thông qua đồ thị trọng số. Bằng cách này, thông tin của toàn bộ cơ sở dữ liệu được sử dụng một cách hiệu quả để tạo thuận lợi cho việc học trong tương lai. Năm 2004, He và các cộng sự đã đề xuất thuật toán *Tra cứu ảnh dựa trên xếp hạng Đa tạp* (MRBIR) để biểu diễn ảnh và mối liên kết của chúng như là một đồ thị. Hệ thống này lan truyền thông tin ảnh đã được gán nhãn thông qua cấu trúc đồ thị của cơ sở dữ liệu ảnh và khai thác sự phân bố các ảnh chưa gắn nhãn để cải thiện độ chính xác tra cứu. Năm 2008, Wang và các cộng sự đã áp dụng thuật toán *phân cụm lan truyền* (APC) làm giảm sự đa dạng đồ thị và bảo tồn cấu trúc đa dạng của nó. Đồ thị này làm lu mờ ảnh hưởng của các ảnh nhiễu trong khi làm nổi bật ảnh hưởng các ảnh tin cậy. Tuy nhiên, hiệu năng tra cứu có thể bị suy giảm khi các cụm không giống với khái niệm ngữ nghĩa. Năm 2007, Lin và các cộng sự đã đề xuất phương pháp *Augmented Relation Embedding-ARE* để biến đổi một không gian ảnh vào trong ngữ nghĩa đa tạp. Bằng việc áp dụng cấu trúc đa tạp ngữ nghĩa này, hệ thống có thể thu được sở thích truy vấn của người dùng. Trong khi đó, một biểu diễn ảnh mới dựa trên đặc trưng được tăng cường cũng được triển khai để thích ứng với phương pháp học *ARE*.

Năm 2013, Wan và cộng sự đã đề xuất phân chia cơ sở dữ liệu ảnh thành các khối có kích cỡ bằng nhau, và sau đó áp dụng thuật toán *MRBIR* trên từng khối. Điểm số tra cứu của từng ảnh là một sự hợp nhất điểm số xếp hạng của tất cả các khối trong ảnh. Cũng vào năm 2013, Liu và các cộng sự đã đề xuất hệ thống xếp hạng đa tạp "*Bidirectional-Isomorphic Manifold Learning*" để thu được nhiều biểu diễn ngữ nghĩa hơn từ ảnh web để khắc phục sự biểu diễn nội dung ngữ nghĩa thiếu chính xác do nhiễu và thông tin dư thừa khía cạnh văn bản và trực quan. Phương

pháp này tối ưu đặc trưng trực quan và không gian văn bản và điều chỉnh hợp nhất trong cả hai không gian với một cấu trúc *tô-pô* được gọi là *ánh xạ đa tạp đảo ngược*. Hệ thống này cũng kết hợp cả chú thích ảnh và phân tích tương quan từ khoá để tăng độ chính xác tra cứu cuối cùng.

Năm 2012, Han và các cộng sự đã đưa ra một *framework* phân lớp ảnh sử dụng các ánh xạ đệ quy cục bộ và toàn cục (*Local and Global Regressive Mapping-LGRM*) trong học đa tạp để học dữ liệu đầu vào và hàm ánh xạ của dữ liệu mẫu đầu ra ở cùng thời điểm. Cuối cùng, nó dự đoán nhãn lớp cho một ảnh đưa vào kiểm tra bằng cách áp dụng phân lớp có giám sát trong đa tạp ít chiều đã được học. Năm 1995, Xu và các cộng sự đã đề xuất chiếu xếp hạng đa tạp thông thường vào trong *framework* tối ưu phân tán *Bregman* bằng cách sử dụng một ma trận *kernel* tối ưu tương đương. Dựa trên phát biểu của họ, hai phương pháp "*hiệu quả*" và "*hiệu quả mở rộng*" được gọi là *DMRE* và *DMRC* được tạo ra để tăng độ chính xác tra cứu và rút ngắn thời gian tính toán.

Tất cả các phương pháp học chuyển đổi trên đạt được độ tra cứu chính xác tốt hơn trong mỗi bước lặp. Tuy nhiên, phương pháp này không áp dụng lịch sử thông tin *RF* được tích luỹ của người dùng để cải thiện đồ thị đa tạp. Phương pháp này cũng không thể chạy trên máy tính khi mà số lượng ảnh trong cơ sở dữ liệu đạt đến một mức độ nhất định do sử dụng một số ma trận vuông lớn. Hơn nữa, tất cả các kỹ thuật học ngắn hạn không thể nắm bắt được ý nghĩa ngữ nghĩa của ảnh và do đó không thể đạt được kết quả tra cứu thoả đáng. Và kỹ thuật này cũng không thể ghi nhớ lịch sử phản hồi của người dùng và do đó không thể sử dụng thông tin đó trong tra cứu tương lai.

2.4 Những thách thức trong phản hồi liên quan

Kỹ thuật Phản hồi liên quan đã đạt được nhiều tiến bộ vượt bậc từ khi nó được giới thiệu vào năm 2007 bởi Liu và các cộng sự. Các phương pháp mới luôn được đưa ra để khắc phục những nhược điểm tồn tại trong nó. Tuy nhiên, với những nhược điểm nguyên thủy của kỹ thuật phản hồi liên quan trong *CBIR* thì đến nay vẫn còn phải được các nhà khoa học nghiên cứu thêm.

Các hạn chế trong phản hồi liên quan của hệ thống *CBIR* như sau:

- Không thể trích chọn ngữ nghĩa mức cao: Hầu hết các kỹ thuật *RF* trong *CBIR* sẽ rất khó để trích chọn ngữ nghĩa mức cao của ảnh khi chỉ có đặc trưng mức thấp được sử dụng trong *RF*. Tuy nhiên, cách này vẫn hoạt động tốt trong việc

tra cứu thông tin văn bản. Bởi vì, việc tra cứu vẫn được dựa trên từ khoá chứ không phải trên các đặc trưng mức thấp.

- Sự khan hiếm và mất cân bằng các mẫu phản hồi: Mỗi người dùng đều không muốn thao tác nhiều hơn số lần lặp phản hồi để có được kết quả tốt nhất. Vì vậy, số lượng mẫu phản hồi gắn nhãn có được từ người dùng trong một phiên *RF* là khá nhỏ so với chiều không gian đặc trưng. Do đó, đối với dữ liệu huấn luyện nhỏ thì hầu hết các thuật toán máy học không thể cho ra kết quả chính xác. Thêm nữa là, số lượng mẫu âm có nhãn thường lớn hơn số lượng mẫu dương có nhãn. Các dữ liệu huấn luyện mất cân đối luôn luôn làm cho việc học phân lớp ít đáng tin cậy hơn. Vì thế, đối với các mẫu dữ liệu huấn luyện nhỏ mà đặc biệt là các mẫu dương thì hiển nhiên sẽ làm giảm độ chính xác của *RF*.
- Xử lý thời gian thực: Quá trình học trong *RF* là trực tuyến và do đó mọi vòng lặp phản hồi bao gồm cả huấn luyện và kiểm tra đều phải thực hiện. Vì thế mà hệ thống sẽ tốn rất nhiều thời gian để xử lý. Có một cách hợp lý để giải quyết vấn đề này là sử dụng phương pháp biểu diễn ảnh và cấu trúc lưu trữ như là một cấu trúc cây phân cấp, v.v…

2.5 CBIR với phản hồi liên quan sử dụng SVM

2.5.1 Support Vector Machine

SVM đã được giới thiệu đầu tiên bởi Vapnik vào cuối những năm 90 và đến nay vẫn còn được quan tâm bởi cộng đồng nghiên cứu học máy [6]. Với nền tảng lý thuyết mạnh mẽ và chặt chẽ, nó đang được sử dụng cho nhiều ứng dụng và là một phương pháp học mẫu nhỏ phổ biến có hiệu năng tốt cho bài toán phân loại mẫu. Giả sử có một tập *n* mẫu được gán nhãn $\mathcal{L} = \{(x_1, y_1), \dots, (x_l, y_l)\}$, với $x_i \in \mathbb{R}^d$ biểu diễn 1 ảnh trong không gian *d* chiều và $y_i \in \{1, -1\}$ là các nhãn. Ý tưởng chính của *SVM* là tìm siêu phẳng

$$f(x) = (w.x) + b \tag{2.1}$$

để chia tách các điểm có y_i*=1* và các điểm có y_i*=-1* sao cho siêu phẳng phân cách có lề cực đại trong khi tỷ lệ lỗi phân lớp là nhỏ nhất. Đây là bài toán quy hoạch toàn phương và nó có thể được giải bởi tìm *w* và *b* sao cho cực tiểu hóa hàm

$$\frac{1}{2} \| \mathrm{w} \|^2 + \mathrm{C} \sum_{\mathrm{i}=1}^{\mathrm{n}} \xi_\mathrm{i} \mathrm{s.t.} \, \mathrm{y_i}(\mathrm{w}.\mathrm{x_i} + \mathrm{b}) \geq 1 - \xi_\mathrm{i}, \xi_\mathrm{i} \geq 0, \mathrm{i} = 1 \dots \mathrm{n}. \tag{2.2}$$

Nếu viết điều kiện phân loại dưới dạng đối ngẫu thì bài toán đối ngẫu của *SVM* là chính là bài toán tối ưu tìm các tham α_i *(i=1..n)* để cực đại hóa hàm

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i . x_j) \quad (2.3)$$

Sao cho $\sum_{i=1}^{n} y_i \alpha_i = 0$, $0 \leq a_i \leq C$, $i = 1 \ldots n$,

Ở đây $K(x_i . x_j)$ là hàm *kernel*. Khi đó hàm phân lớp *SVM* sẽ là

$$f(x) = sign\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i . x_j) + b\right) \quad (2.4)$$

và đường bao quyết định sẽ là: $\sum_{i=1}^{l} \alpha_i y_i K(x_i . x_j) + b = 0$.

2.5.2 Phản hồi bị động và học chủ động

Trong tra cứu ảnh với phản hồi liên quan dựa trên *SVM*, đường bao quyết định đã được sử dụng để đo sự liên quan giữa ảnh truy vấn và các mẫu đưa vào. Nói chung, các mẫu có giá trị tuyệt đối của hàm phân lớp càng lớn thì khả năng tin cậy dự đoán sẽ cao. Trong phương pháp phản hồi liên quan dựa trên *SVM* truyền thống, người dùng sẽ chọn các mẫu được xếp hạng trên cùng, tức là các mẫu có giá trị lớn nhất của hàm *SVM f(x)* cho việc huấn luyện *SVM*.

$$x^* = argmax_{x \in \mathcal{U}}(f(x)) \quad (2.5)$$

Chiến lược này được gọi là phản hồi bị động. Nó hướng tới việc chọn các mẫu liên quan nhất. Tuy nhiên, những mẫu này có thể không phải là các mẫu thông tin nhất cho huấn luyện *SVM*, do vậy đường bao quyết định của *SVM* có thể sẽ không được cải thiện.

Ngược lại với phương pháp này là học *SVM* chủ động đang được quan tâm và có nhiều hứa hẹn trong những năm gần đây. Học chủ động hướng tới việc chọn các mẫu không chắc chắn, đó là các mẫu gần nhất với đường bao quyết định *SVM*:

$$x^* = argmin_{x \in \mathcal{U}}|f(x)| \quad (2.6)$$

Với phương pháp này, đường bao quyết định có thể sẽ được cập nhật tốt hơn. Tuy nhiên việc tìm ra các mẫu thông tin nhất cho huấn luyện *SVM* là một thách thức.

CHƯƠNG 3: ÁP DỤNG CÀI ĐẶT THỬ NGHIỆM

3.1 Cài đặt

3.1.1 Cơ sở dữ liệu

Chương trình được cài đặt trên Microsoft Visual Studio 2010 với ngôn ngữ C#. Và một thư viện liên kết *LibSVMsharp.dll* hỗ trợ cho việc áp dụng thuật toán *SVM* để huấn luyện.

Tập cơ sở dữ liệu thử nghiệm bao gồm 10800 ảnh lấy từ cơ sở dữ liệu ảnh *COREL*. Cơ sở dữ liệu bao gồm 80 lớp khác nhau, mỗi lớp khoảng 100 ảnh.

3.1.2 Trích chọn đặc trưng và đối sánh

Để trích chọn các đặc trưng của ảnh chương trình sử dụng thư viện *FELib.dll*. Và sau đó, chương trình đã sử dụng 5 loại đặc trưng dưới đây để để biểu diễn cho một ảnh và chúng được biểu diễn bởi một *vector* 809 phần tử:

- *Color histogram, color moments* từ phần tử từ 1 đến 81
- *Edge histogram* từ 82 đến 118
- *Gabor wavelets transform*: các phần tử từ 119 đến 238
- *Local Binary Pattern*: các phần tử từ 239 đến 297
- *GIST*: các phần tử từ 297 đến 809

Để đo độ tương tự giữa hai ảnh, chương trình sử dụng độ đo *Euclid.*

Để huấn luyện *SVM*, chương trình sử dụng các tham số mặc định của thư viện *LibSVMsharp.dll*

3.2 Các chức năng chương trình

3.2.1 Mở ảnh truy vấn

- Mở ảnh bằng cách chọn *File -> Open* trên *Menu* chức năng.
- Trích chọn đặc trưng ảnh truy vấn.
- Hiển thị ảnh truy vấn lên *picturebox*.

3.2.2 Tra cứu ảnh

- Tính toán độ tương tự giữa ảnh truy vấn với từng ảnh trong cơ sở dữ liệu ảnh thông qua hàm tính toán khoảng cách *Euclid*.
- Sắp xếp các ảnh giảm dần theo đo độ tương tự.
- Hiển thị các ảnh đã được sắp xếp lên khung *retrieval results*.

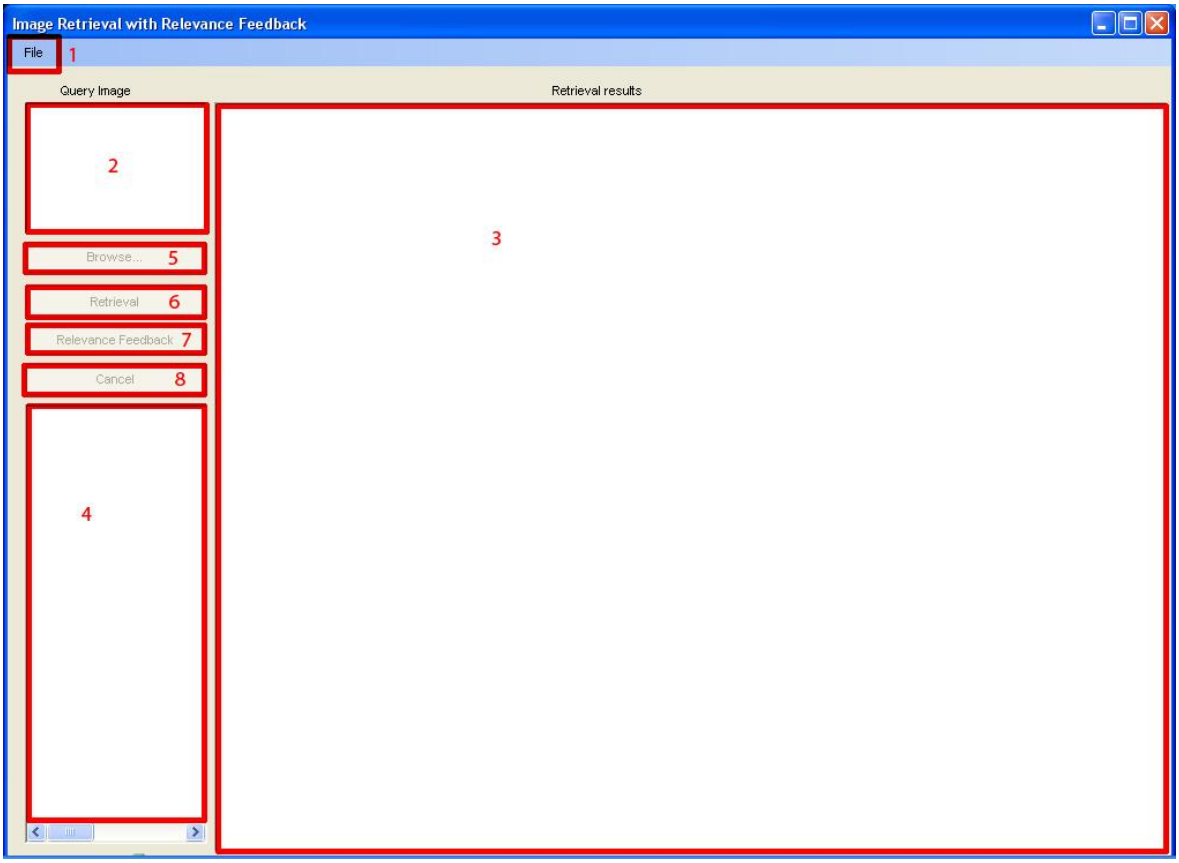
- Chọn một số ảnh gần nhất với ảnh truy vấn hiển thị lên khung *relevance feedback* để người dùng gán nhãn.
- Người dùng gán nhãn có liên quan (*+1*) cho ảnh gần với ảnh truy vấn nhất bằng cách kích vào ảnh đó, và những ảnh còn lại không được người dùng kích thì được gán nhãn là không liên quan (*-1*).

3.2.3 Phản hồi liên quan

- Các ảnh đã gán nhãn sẽ được dùng để huấn luyện *SVM* để tìm ra đường bao quyết định phân lớp.
- Chương trình tính toán khoảng cách từ các ảnh chưa được gán nhãn trong cơ sở dữ liệu đến đường bao quyết định *SVM* (khoảng cách này được gọi là *disSVM*).
- Hiển thị các ảnh có giá trị *disSVM* lớn nhất lên khung *retrieval results* (tức là sắp xếp các ảnh chưa được gán nhãn giảm dần theo giá trị *disSVM*).
- Chọn các ảnh gần với đường bao quyết định *SVM* để hiển thị lên khung *relevance feedback* (tức là sắp xếp các ảnh chưa được gán nhãn tăng dần theo giá trị tuyệt đối của *disSVM*).
- Người dùng sẽ thực hiện tiếp việc gán nhãn cho các ảnh này.
- Quá trình này được lặp đi lặp lại cho đến khi người dùng hài lòng với kết quả tra cứu.

3.3 Kết quả thử nghiệm

3.3.1 Giao diện chương trình



Hình 3-1: Giao diện chương trình

Các thành phần chính của chương trình bao gồm:

1. *Menu* chức năng.

2. Một *picture box* để hiển thị ảnh truy vấn mẫu.

3. Một khung *retrieval results* để hiển thị kết quả các ảnh tra cứu.

4. Một khung *relevance feedback* để hiển thị các ảnh liên quan để người dùng lựa chọn.

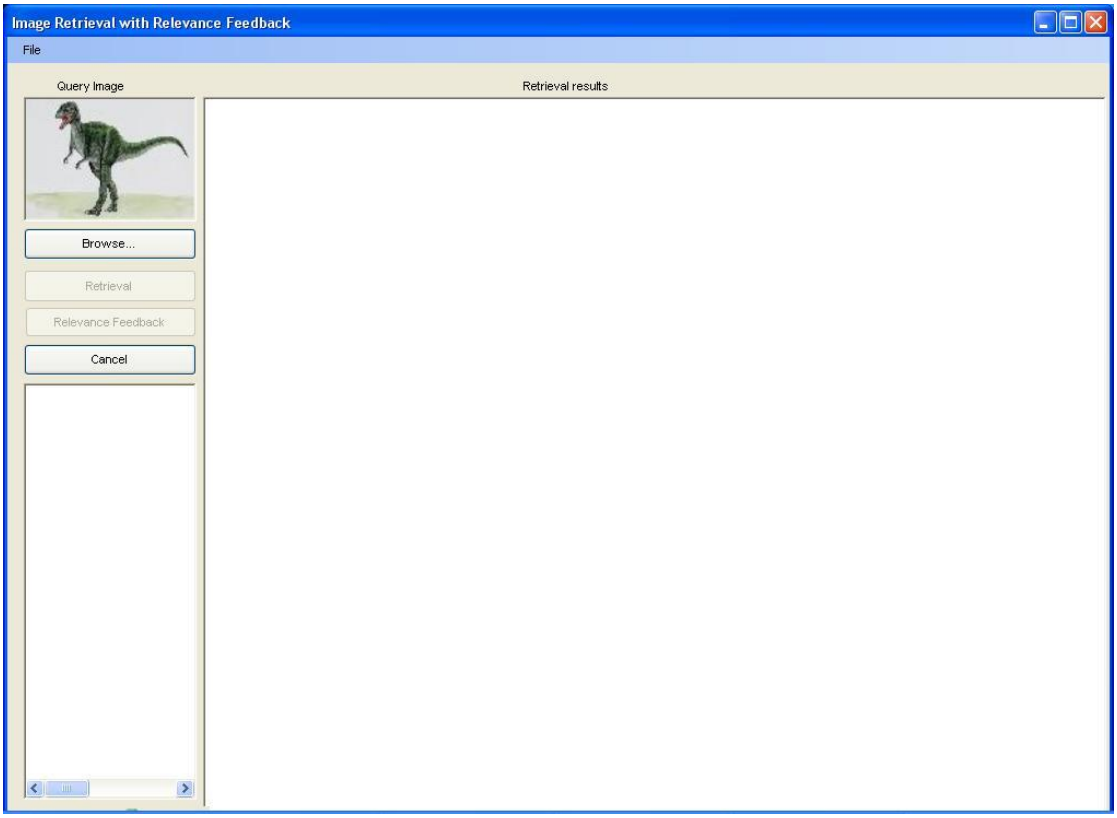
5. Nút "*Browse*" để chọn thư mục chứa cơ sở dữ liệu ảnh.

6. Nút "*Retrieval*" để tra cứu ảnh theo độ đo tương tự của đặc trưng mức thấp.

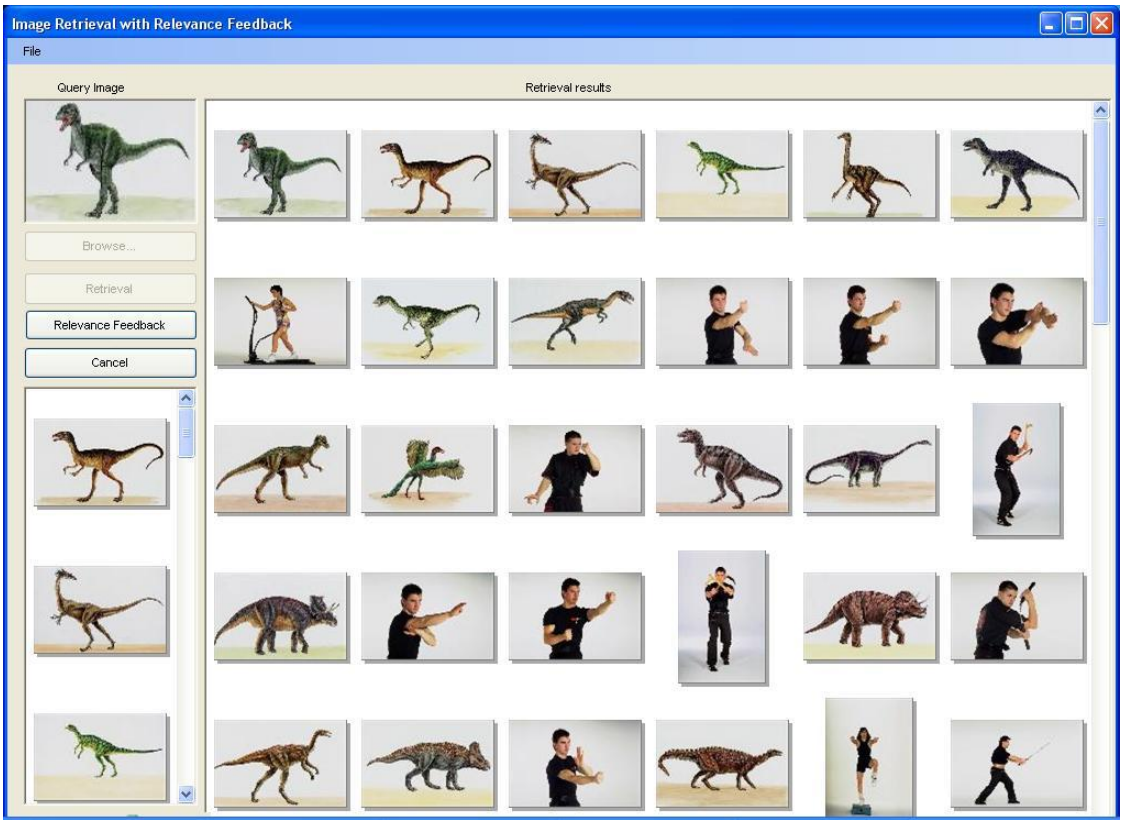
7. Nút "*Relevance Feedback*" để thực hiện phản hồi liên quan.

8. Nút "*Cancel*" để dừng quá trình.

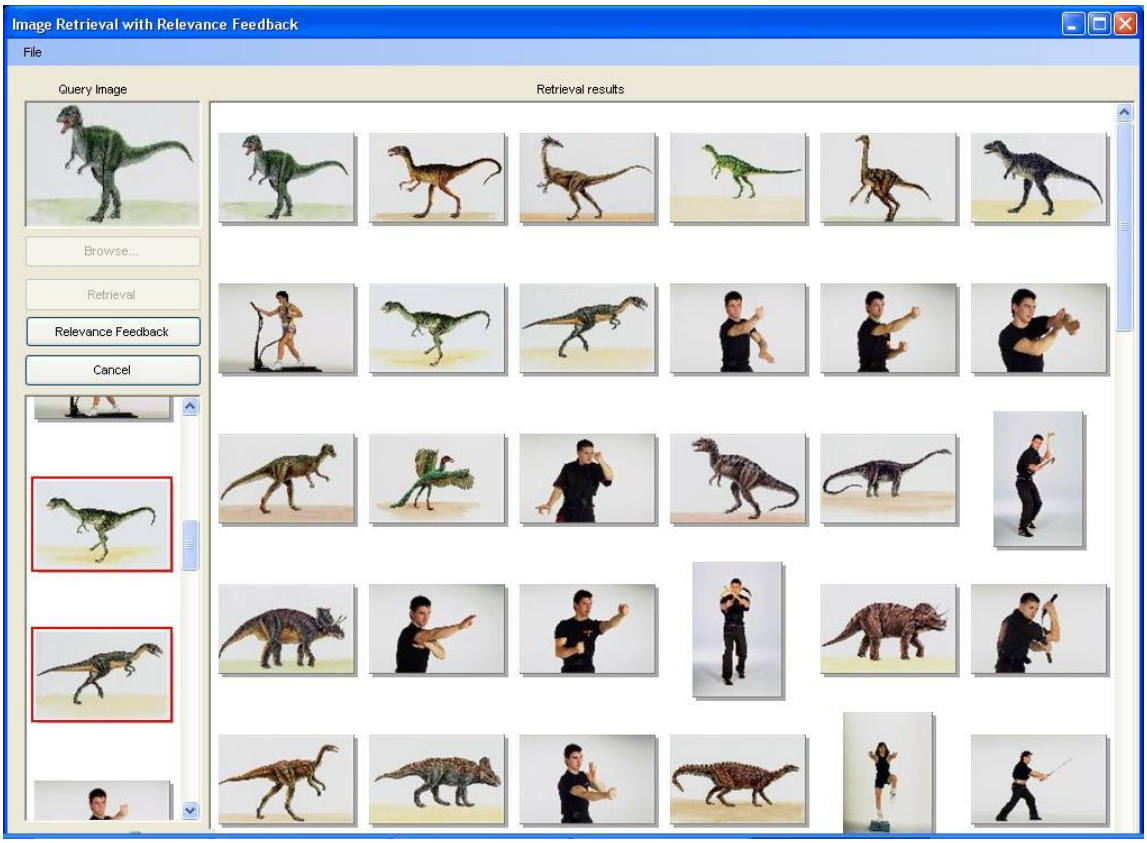
3.3.2 Một số kết quả thử nghiệm



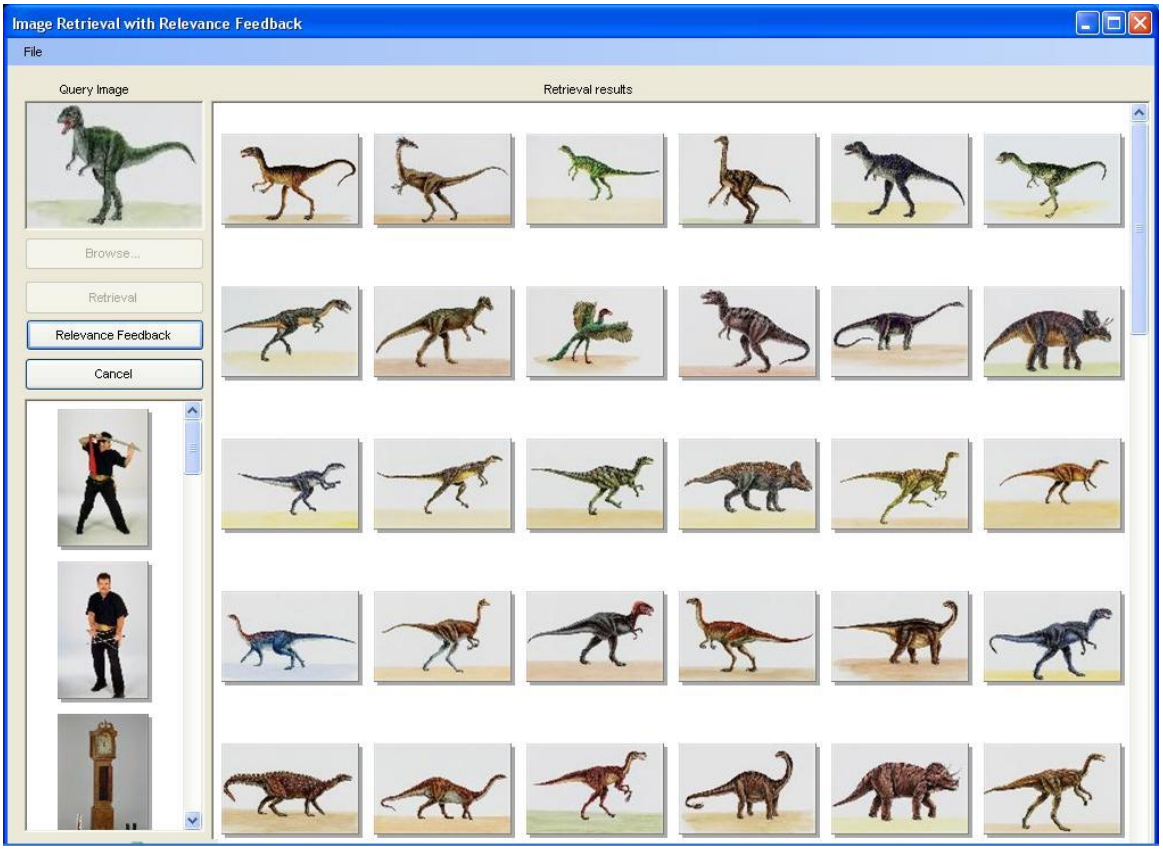
Hình 3-2: Người dùng chọn ảnh truy vấn



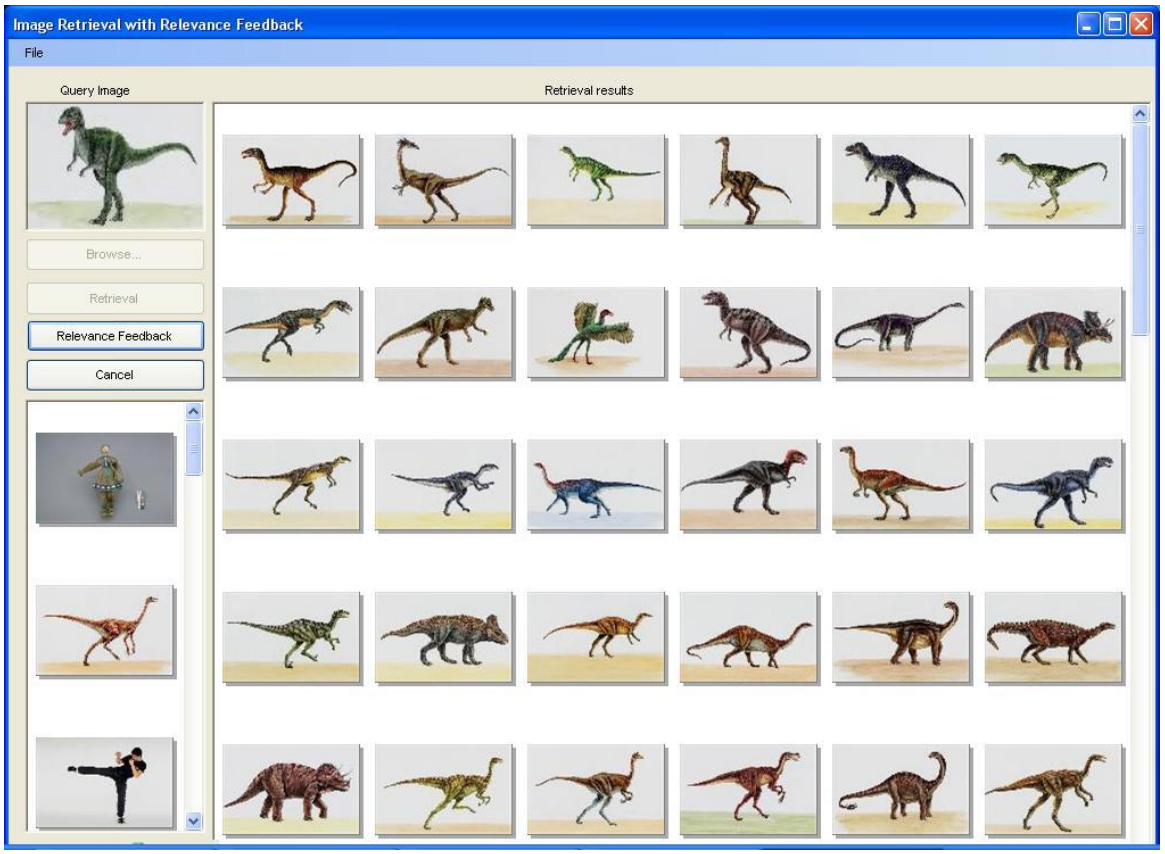
Hình 3-3: Kết quả tra cứu ban đầu



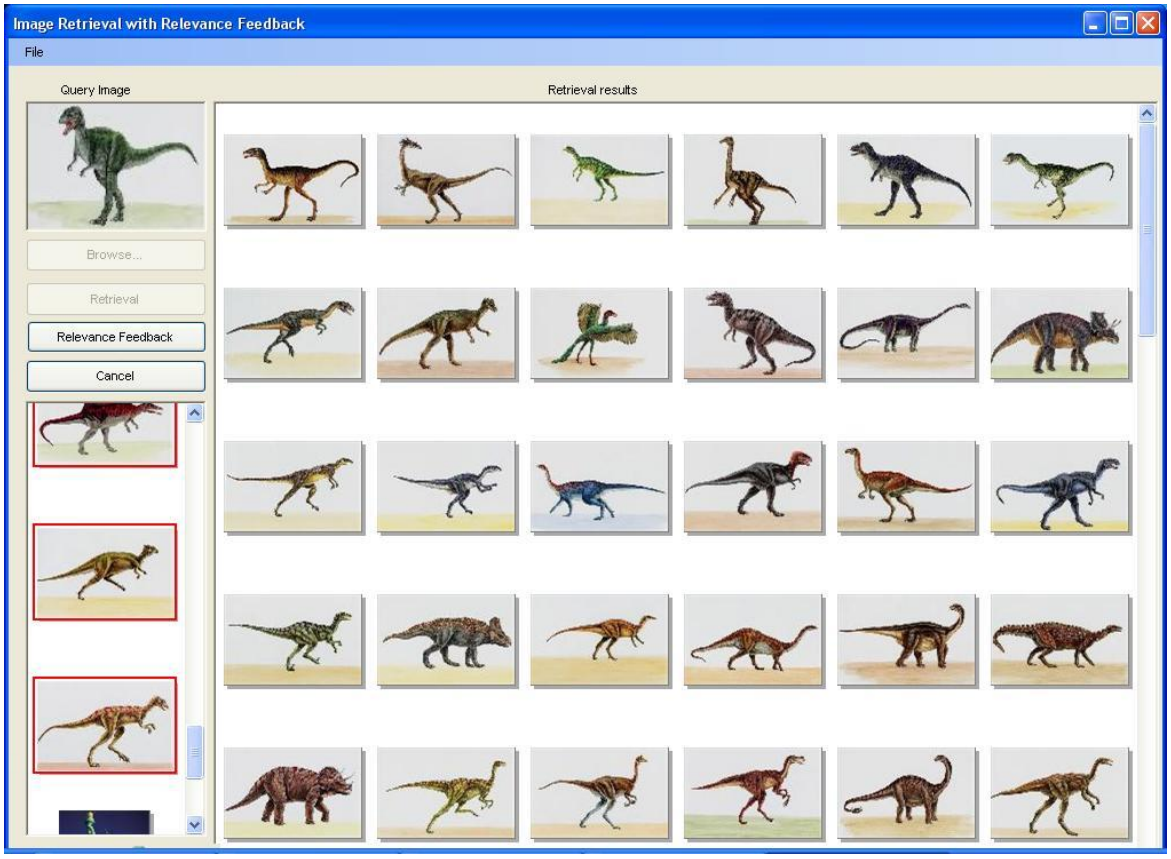
Hình 3-4: Người dùng chọn ảnh liên quan lần 1



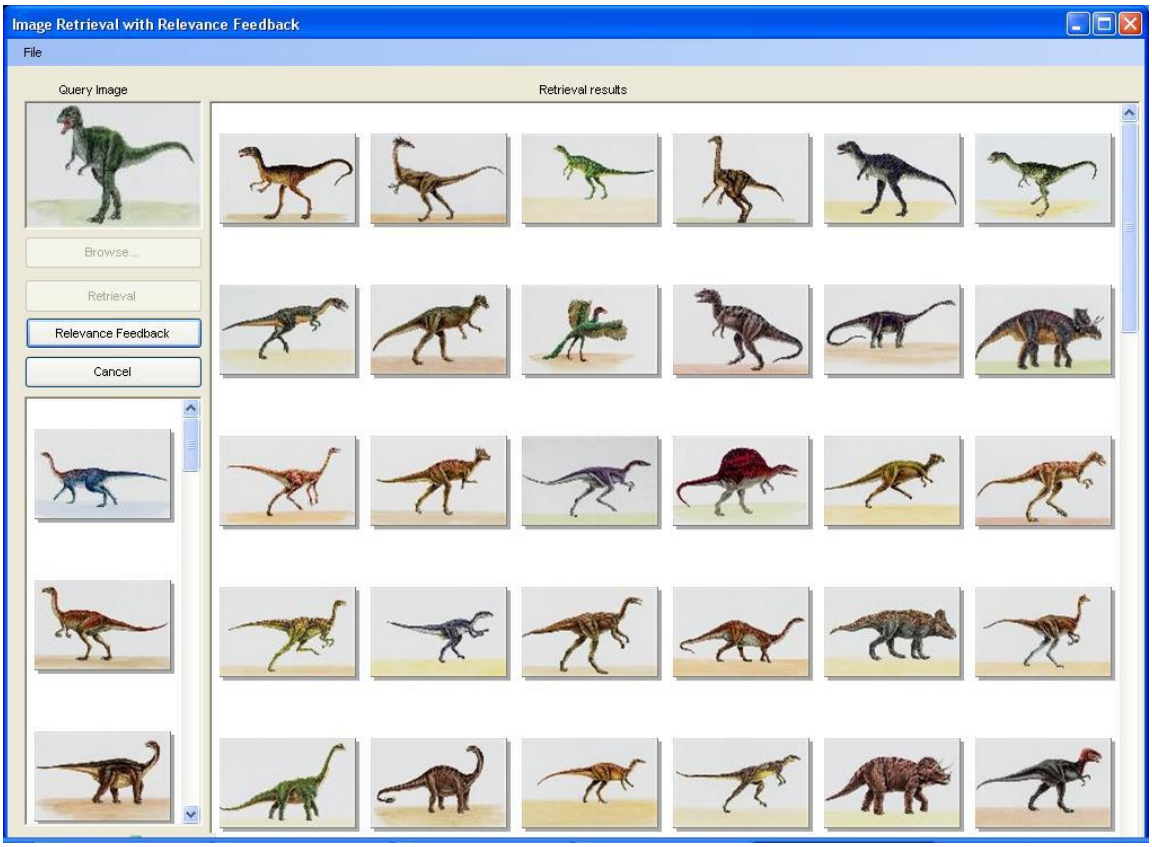
Hình 3-5: Kết quả sau vòng lặp phản hồi thứ nhất



Hình 3-6: Kết quả sau vòng lặp phản hồi thứ hai



Hình 3-7: Người dùng chọn ảnh liên quan lần 3



Hình 3-8: Kết quả sau vòng lặp phản hồi thứ ba

3.4 Một số nhận xét về chương trình

Có thể thấy rằng, việc áp dụng kỹ thuật máy học *SVM* vào phản hồi liên quan đã làm tăng độ chính xác tra cứu ảnh dựa theo nội dung. Tuy nhiên, vẫn tồn tại một số mặt hạn chế của phương pháp này.

Các mặt hạn chế là:

- Máy tính mất nhiều thời gian để tính toán.
- Thiếu lịch sử thông tin phản hồi của người dùng để hệ thống có thể học ý định của người dùng trong dài hạn.
- Chương trình chỉ áp dụng duy nhất phương pháp máy học *SVM*, mà chưa áp dụng kết hợp các thuật toán máy học khác nhằm nâng cao hiệu năng tra cứu ảnh.

KẾT LUẬN

Sau một thời gian tìm hiểu và nghiên cứu đồ án đã đạt được một số kết quả sau:

- Tìm hiểu được cấu trúc của một hệ thống tra cứu ảnh dựa trên nội dung.
- Tìm hiểu được một số phương pháp làm giảm khoảng cách ngữ nghĩa trong tra cứu ảnh dựa trên nội dung.
- Tìm hiểu các phương pháp trong phản hồi liên quan.
- Tìm hiểu thuật toán máy học SVM được áp dụng vào trong hệ thống phản hồi liên quan để học phản hồi của người dùng.
- Xây dựng được chương trình thử nghiệm.

Tuy nhiên do thời gian có hạn trong quá trình nghiên cứu đề tài nên vẫn chỉ dừng lại ở việc xây dựng chương trình thử nghiệm. Ngoài ra, chương trình mới chỉ dừng lại ở mức áp dụng một thuật toán máy học SVM cho phản hồi liên quan, chưa áp dụng được các thuật toán máy học khác vào trong chương trình thử nghiệm để so sánh hiệu năng tra cứu của từng thuật toán với nhau.

Do đó, em rất mong nhận được sự đóng góp ý kiến từ các Thầy Cô và các bạn để em có thêm kiến thức và kinh nghiệm tiếp tục hoàn thiện nội dung nghiên cứu trong đề tài.

Em xin chân thành cảm ơn!

TÀI LIỆU THAM KHẢO

Tài liệu Tiếng Việt:

[1] Nguyễn Thị Hoàn, "*Phương pháp trích chọn đặc trưng ảnh trong thuật toán học máy tìm kiếm ảnh áp dụng vào bài toán tìm kiếm sản phẩm.*" Khoá luận tốt nghiệp, Đại học Công Nghệ, năm 2010.

Tài liệu Tiếng Anh:

[2] Khuat Thi Thu Ha, project: "*Content-based image retrieval with relevance feedback*", Final report master in information and communication and technology, University of Science and Technology of Hanoi, September 2013.

[3] Chang, Ran, "*Effective graph-based content-based image retrieval systems for large-scale and small-scale image databases*", Doctor of Philosophy, Utah State University 2013.

[4] Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-ying Ma, "*A survey of content-based image retrieval with high-level semantics*," Pattern recognition, volume 40, issue 1, January, 2007, 262-282.

[5] Dr. Fuhui Long, Dr. Hongjiang Zhang and Prof. David Dagan Feng, "*Fundamentals of content-based image retrieval*", International journal of computer science and information technologies, vol.3 (1), 2012, 3260 – 3263.

[6] Ngo Truong Giang, Khuat Thi Thu Ha, Ngo Quoc Tao and Nguyen Duc Dung, "*Interactive Image Retrieval with Active Support Vector Machine Learning*", Department of Information Technology, HaiPhong Private University, Institute of Information Technology, Vietnamese Academy of Sciences and Technology, University of Science and Technology of Hanoi. FAIR - Thai Nguyen, 20-21/6/2014.