# CẢI TIẾN ĐỘ CHÍNH XÁC TRA CỨU ẢNH THÔNG QUA HỌC SÂU VÀ HỌC ĐỘ ĐO KHOẢNG CÁCH TỐI ƯU

**[*]Đào Thị Thúy Quỳnh**

[*] Khoa Công nghệ thông tin 1, Học Viện Công Nghệ Bưu Chính Viễn Thông

***Tóm tắt-*** Tra cứu ảnh dựa vào nội dung được thực hiện bởi việc so sánh độ đo tương tự giữa biểu diễn ảnh truy vấn và biểu diễn cơ sở dữ liệu ảnh. Do đó, hiệu quả của phương pháp tra cứu ảnh bị ảnh hưởng rất nhiều bởi biểu diễn ảnh và độ đo tương tự. Gần đây, học sâu được sử dụng và đem lại hiệu quả cao trong các bài toán phân lớp, nhận dạng ảnh, các đặc trưng ảnh được học bởi mô hình CNN mang tính ngữ nghĩa cao. Trong bài báo này, chúng tôi sẽ đề xuất phương pháp tra cứu ảnh IRDLoM (Image Retrieval using Deep learning and optimal distance metric) sử dụng mạng CNN để xây dựng bộ đặc trưng và tìm một phép chiếu tuyến tính với một độ đo tương tự cải tiến. Phần thực nghiệm cung cấp các kết quả thực nghiệm để minh chứng độ chính xác của phương pháp đề xuất.

***Từ khóa:*** Content-based image retrieval, deep learning, similarity measures, mahalanobis metric distance.

## I. MỞ ĐẦU

Tra cứu ảnh dựa vào nội dung (CBIR-Content Based Image Retrieval) đã nhận được nhiều sự quan tâm trong thập kỷ qua, do nhu cầu xử lý hiệu quả lượng dữ liệu đa phương tiện khổng lồ và tăng nhanh chóng. Nhiều hệ thống CBIR đã được phát triển, gồm QBIC [21], Photobook [22], MARS [23], PicHunter [24] , Blobworld [25].

Trong một hệ thống CBIR tiêu biểu, các đặc trưng ảnh trực quan mức thấp (màu, kết cấu và hình dạng) được trích rút tự động và biểu diễn thành các véc tơ đặc trưng tương ứng cho mục tiêu mô tả ảnh và so sánh độ tương tự. Để tìm kiếm các ảnh mong muốn, người dùng đưa một ảnh làm mẫu truy vấn và hệ thống trả lại một tập các ảnh tương tự dựa vào các đặc trưng được trích rút. Khi các hệ thống trình bày một tập các ảnh được xem là tương tự đối với truy vấn, người dùng có thể lấy ra những ảnh liên quan nhất với ảnh truy vấn được cho, và hệ thống điều chỉnh truy vấn sử dụng chúng. Phản hồi liên quan dựa vào các kỹ thuật CBIR không yêu cầu người dùng cung cấp các truy vấn khởi tạo chính xác, nhưng đánh giá truy vấn lý tưởng của người dùng bằng sử dụng các ảnh liên quan phản hồi bởi người dùng.

Tác giả liên hệ: Đào Thị Thúy Quỳnh
Email: quynhdao.ptit@gmail.com
Đến tòa soạn: 8/2020, chỉnh sửa: 9/2020, chấp nhận đăng: 10/2020.

Do đó, biểu diễn ảnh bởi véc tơ đặc trưng và độ đo tương tự là hai yếu tố chính ảnh hưởng tới hiệu quả của hệ thống CBIR. Nâng cao hiệu quả của hệ thống CBIR là một vấn đề thách thức trong nghiên cứu. Để nâng cao hiệu quả, chúng ta cần giảm khoảng trống ngữ nghĩa trong CBIR, khoảng trống ngữ nghĩa thể hiện những hạn chế của biểu diễn ảnh bởi đặc trưng mức thấp được trích rút tự động và ngữ nghĩa của bức ảnh do con người cảm nhận. Để giảm khoảng trống ngữ nghĩa này, đã có một số đề xuất đưa các kỹ thuật học máy vào trong quá trình tra cứu ảnh. Những năm gần đây, học sâu đã nâng cao được hiệu quả của các bài toán nhận dạng, phân loại đối tượng. Với mong muốn nâng cao hiệu quả ngay từ quá trình xây dựng bộ đặc trưng biểu diễn ảnh, phương pháp đề xuất sẽ sử dụng cấu trúc mạng CNN để xây dựng bộ đặc trưng có tính ngữ nghĩa cao. Bên cạnh đó, phương pháp đề xuất sẽ kết hợp kỹ thuật phân lớp LDA và học độ đo tương tự (Learning similarity measures) để đưa một độ đo tương tự cải tiến phù hợp hơn với dữ liệu.

Ý tưởng của học độ đo khoảng cách là tìm một độ đo khoảng cách tối ưu mà tối thiểu được khoảng cách giữa các cặp ảnh tương tự nhau và tối đa hóa khoảng cách giữa những cặp ảnh không tương tự. Sau đó, độ đo khoảng cách tối ưu này sẽ được dùng để phân hạng lại toàn bộ tập ảnh và trả về kết quả. Chúng tôi đề xuất một kỹ thuật hiệu cứu ảnh hiệu quả, kỹ thuật có tên là IRDLoM (Image Retrieval using Deep learning and optimal distance metric). Bằng thực nghiệm trên cơ sở dữ liệu ảnh gồm 10.800 ảnh , chúng tôi sẽ chỉ ra sự chính xác của phương pháp đề xuất.

Phần còn lại của bài báo được tổ chức như sau. Trong Phần 2, trình bày chi tiết phương pháp đề xuất. Phần 3 mô tả các thực nghiệm hiệu năng của chúng tôi và thảo luận các kết quả. Cuối cùng, chúng tôi đưa ra kết luận.

## II. NGHIÊN CỨU LIÊN QUAN

Tra cứu ảnh dựa vào nội dung sử dụng học khoảng cách đã nhận được sự quan tâm trong cộng đồng nghiên cứu [6, 9, 13, 14, 15, 16, 17,18]. Dữ liệu đầu vào của các thuật toán học khoảng cách trong tra cứu ảnh thường được chia làm hai nhóm: (1) chỉ xem xét đến các cặp ảnh tương tự và (2) xem xét cả các cặp ảnh tương tự và các cặp ảnh không tương tự.

Ý tưởng điều chỉnh trọng số của hàm khoảng cách đã được áp dụng vào các hệ thống tra cứu ảnh, chẳng hạn như phương pháp SRIR [19]. Phương pháp này thường tận dụng thông tin của tập ảnh tương tự, xem xét tới sự phân tán của dữ liệu trên mỗi chiều và biểu diễn bởi một ma

trận đường chéo. Từ đó đưa ra một hàm khoảng cách Euclid cải tiến và áp dụng nó vào phân hạng toàn bộ tập ảnh.

Phương pháp MCML [4], các phương pháp này học một độ đo khoảng cách Mahalanobis sao cho các mẫu cùng một lớp sẽ được ánh xạ tới cùng một điểm. Bài toán học độ đo khoảng cách được đưa về bài toán tối ưu lồi và tìm nghiệm theo phương pháp Gradient-descent. Tuy nhiên, việc tìm nghiệm của bài toán tối ưu bởi phương pháp Gradient-descent có chi phí tính toán lớn.

Phương pháp LMNN [5] với ý tưởng cực tiểu khoảng cách các mẫu cùng nhãn nằm trong lân cận k-NN và cực đại khoảng cách các mẫu khác nhãn bởi một lề lớn hơn mà sử dụng hàm khoảng cách Mahalanobis. Ý tưởng này được mô hình hóa bởi một bài toán tối ưu và giải quyết nó bởi phương pháp SDP [3] từ đó tìm ra độ đo khoảng cách cải tiến.

Thuật toán học trực tuyến cho độ tương tự ảnh cỡ lớn (OASIS) [18] được thiết kế chuyên biệt để làm việc với các ràng buộc cặp. Tuy nhiên, chúng dựa trên các giả thiết mạnh về dữ liệu đầu vào hoặc cấu trúc của các ràng buộc (yêu cầu dữ liệu đầu vào là các véc tơ thưa). Do đó, nó khó có thể áp dụng được trong thực tế.

Phương pháp Xing [20] với ý tưởng đưa về bài toán tối ưu dạng lồi mà cực tiểu hóa tổng khoảng cách của các cặp ảnh tương tự với ràng buộc tổng khoảng cách các cặp ảnh không tương tự đạt cực đại. Ở pha khởi tạo, phương pháp sử dụng hàm khoảng cách Euclid cải tiến với A=I. Sau đó, phương pháp Xing đưa ra một hàm khoảng cách cải tiến với A là kết quả của bài toán tối ưu lồi nói trên. Tuy nhiên, phương pháp của Xing cũng có chi phí tính toán lớn do sử dụng phương pháp giải Gradient-descent để tìm nghiệm và cũng chưa khai thác tập ảnh tương tự một cách hiệu quả.

Với phương pháp RCA [8], ý tưởng của phương pháp này chỉ sử dụng các cặp ảnh tương tự, tìm một phép biến đổi dữ liệu dựa vào ma trận phương sai sinh ra từ tập ảnh tương tự. Từ đó, cải tiến hàm khoảng cách Mahalanobis bằng cách thay đổi ma trận trọng số. Mặc dù, phương pháp RCA này có chi phí tính toán hiệu quả hơn phương pháp của Xing nhưng phương pháp RCA chỉ xem xét tới tập ảnh tương tự.

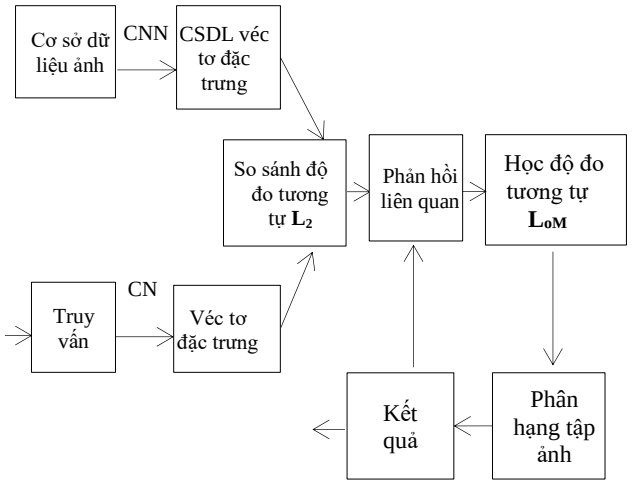
Từ phân tích ưu điểm và hạn chế của những nghiên cứu liên quan ở trên, chúng tôi đề xuất phương pháp tra cứu ảnh với hàm khoảng cách cải tiến. Việc cải tiến hàm khoảng cách dựa trên việc cực đại hóa thương giữa tổng khoảng các cặp ảnh không tương tự và tổng khoảng cách các cặp ảnh tương tự. Trong ý tưởng này, chúng ta xem xét được cả tập ảnh tương tự và không tương tự để tìm được ma trận trọng số và cải tiến hiệu quả của phương pháp tra cứu.

## III. PHƯƠNG PHÁP TRA CỨU ẢNH ĐỀ XUẤT

Phương pháp đề xuất sẽ thực hiện xây dựng bộ đặc trưng dựa vào học sâu, từ k-NN sẽ trả về tập ảnh khởi tạo cho người dùng. Quá trình phản hồi liên quan được thực hiện, người dùng sẽ lựa chọn ra tập ảnh phù hợp với mong muốn là tập mẫu liên quan. Lấy được tập mẫu liên quan, phương pháp sẽ thực hiện huấn luyện để tìm ra một phép chiếu tuyến tính thỏa mãn phương sai giữa các mẫu cùng tập liên quan là cực tiểu và cực đại hóa phương sai giữa mẫu liên quan và không liên quan. Sau đó, phương pháp sẽ thực hiện xây dựng một độ đo tương tự cải tiến Mahalanobis bằng thực hiện tìm ma trận tối ưu M trong công thức độ đo tương tự cải tiến.

### *A. Tổng quan phương pháp*

Phương pháp tra cứu ảnh đề xuất IRDLoM được mô tả trên Hình 1. Phương pháp sẽ sử dụng mô hình CNN đã được huấn luyện trên một tập dữ liệu, sau đó sử dụng cấu trúc mạng làm khởi tạo để trích rút đặc trưng mức cao, đó là quá trình biểu diễn ảnh bởi véc tơ đặc trưng. Khi người dùng đưa vào một ảnh truy vấn, phương pháp cũng thực hiện trích rút đặc trưng tương tự như thực hiện với ảnh cơ sở dữ liệu. Phương pháp sẽ thực hiện so sánh độ tương tự giữa véc tơ đặc trưng ảnh truy vấn và tập véc tơ đặc trưng của cơ sở dữ liệu ảnh sử dụng độ đo Euclid và trả về tập ảnh kết quả khởi tạo cho người dùng. Người dùng sẽ thực hiện quá trình phản hồi liên quan, lựa chọn ra những ảnh phù hợp với mong muốn. Tiếp theo, thông tin phản hồi bao gồm tập ảnh liên quan và không liên quan được đưa vào học độ đo khoảng cách và tối ưu hóa trọng số của hàm khoảng cách cải tiến. Sau đó, tất cả các ảnh được sắp xếp lại dựa trên giá trị của hàm khoảng cách Mahalanobis cải tiến. Nếu người dùng chưa thỏa mãn với các kết quả, quá trình phản hồi liên quan sẽ được lặp lại để trả về tập ảnh kết quả cho người dùng.




*Hình 1. Sơ đồ của phương pháp đề xuất.*

### *B. Biểu diễn ảnh sử dụng học sâu*

Trong những năm gần đây, mạng CNN đã đem lại hiệu quả tuyệt vời trong trong lĩnh vực thị giác máy như bài toán phân lớp ảnh, xác định đối tượng, phân đoạn ngữ nghĩa. Từ đó, cũng có nhiều nghiên cứu về tra cứu ảnh dựa vào nội dung (CBIR) sử dụng CNN và nhận được kết quả tốt.

Trong [7] chỉ ra một số cách tiếp cận để cải tiến hiệu quả của hệ thống CBIR sử dụng học sâu trong việc xây dựng ra bộ đặc trưng có tính ngữ nghĩa cao hơn: (1) sử dụng một mô hình CNN đã tiền huấn luyện và xây dựng ra bộ đặc trưng ảnh dùng khoảng cách $L_2$ để so sánh độ đo tương ứng giữa các véc tơ đặc trưng; (2) vẫn dùng mô hình CNN đã được tiền huấn luyện để xây dựng ra bộ đặc trưng, tuy nhiên nó cải tiến bằng cách sử dụng học độ đo khoảng cách (DML) để có được một độ đo tương tự thích hợp với dữ liệu hơn ở pha so sánh độ tương tự; và (3) với một bộ dữ liệu cụ thể nào đó, huấn luyện lại mô hình CNN kết hợp với một bộ phân lớp cụ thể, sau đó sử dụng

độ đo như cách tiếp cận (1) hoặc (2) là hoàn thiện một phương pháp tra cứu ảnh sử dụng học sâu.

Trong [7] đã giới thiệu cách tiếp cận (1) đó cũng là một trường hợp cải tiến của cách tiếp cận (2). Giả sử, chúng ta có hai ảnh trong CSDL là $I_i$ và $I_j$, quá trình trích rút đặc trưng sử dụng một mô hình CNN đã được tiền huấn luyện trên tập dữ liệu lớn, sau đó sử dụng mô hình làm khởi tạo để trích rút đặc trưng mức cao. Quá trình này còn được gọi là quá trình học biểu diễn ảnh, tương ứng bộ đặc trưng mức cao là $x_i$ và $x_j$. Độ đo tương tự được dùng để so sánh giữa hai đặc trưng này là $L_2$:

$$similarity(x_i, x_j) = \|x_i - x_j\|_2 \quad (1)$$

Công thức (1) thể hiện độ tương tự giữa ảnh $I_i$ và $I_j$, độ tương tự càng lớn thì ảnh $I_i$ và $I_j$ càng tương tự nhau.

Độ đo tương tự theo cách tiếp cận thứ (2) để so sánh giữa hai véc tơ đặc trưng của ảnh được tính bởi công thức $L_A$:

$$similarity(x_i, x_j) = \|x_i - x_j\|_A = (x_i - x_j)^T A (x_i - x_j) \quad (2)$$

Với ma trận A được học từ quá trình học độ đo tương tự với điều kiện M là ma trận xác định dương, vì độ tương tự phải dương, và độ tương tự đạt giá trị nhỏ nhất khi $x_i = x_j$. Độ đo tương tự trong cách tiếp cận này sẽ là cách tiếp cận (1) khi ma trận A là một ma trận đơn vị $A = I$. Một cách khác, đó chính là trường hợp đặc biệt khi chúng ta xem xét đến sự tương quan giữa các thành phần đặc trưng trong cách tiếp cận (1). Hơn thế nữa, mỗi thành phần đặc trưng lại có độ tương tự khác nhau nên thường độ đo tương tự ở cách tiếp cận (2) đem lại hiệu quả hơn.

Phương pháp đề xuất sẽ thực hiện xây dựng bộ đặc trưng dựa vào học sâu, từ k-NN lấy được, phương pháp sẽ thực hiện huấn luyện với mô hình LDA. Sau đó, phương pháp sẽ xây dựng một độ đo tương tự cải tiến bằng cách tận dụng tập mẫu dương lấy ý tưởng từ cách tiếp cận (2) để xây dựng nên ma trận A trong công thức độ đo tương tự (2), ma trận M là một ma trận đầy đủ nó sẽ phản ánh được sự tương quan của dữ liệu trên từng đặc trưng và giữa các đặc trưng.

Thuật toán học biểu diễn đặc trưng ảnh **(Representation image learning)** dưới đây thực hiện học biểu diễn ảnh dựa vào tiền huấn luyện mạng học sâu CNN thu được tập đặc trưng mức cao. Thuật toán nhận đầu vào là một tập các ảnh và mô hình đã tiền huấn luyện CNN trên bộ ảnh ImageNet.

**Thuật toán 1.1. Thuật toán RIL (Representation image learning)**

**Input: -** Tập các ảnh X = $\{x_1, x_2, \ldots, x_n\}$ với $x_i \in R^m$

- Mô hình tiền huấn luyện M

**Output: -** Tập biểu diễn ảnh S = $\{s_1, s_2, \ldots, s_n\}$ với $s_i \in R^d$

**1.** Model ← LoadModel(M);

**2.** $S \leftarrow \emptyset$

**3.** for i = 1,…,n do

3.1. $s_i \leftarrow$ ExtractFeature($x_i$,Model);

3.2. $S \leftarrow S \cup s_i$

**4.** Return S

*C. Một độ đo khoảng cách cải tiến*

Cho đến nay, cũng có một số cách tiếp cận học khoảng cách khác nhau mà khai thác tính chất của tập phản hồi từ phía người dùng trong quá trình tra cứu ảnh. Tuy nhiên, các phương pháp đã có thường chỉ xem xét tới tập mẫu dương (positive samples) mà chưa xem xét tới tập mẫu âm. Ý tưởng cơ bản của phân tích thành phần phân biệt (**DCA-**Discriminative Component analysis) là tìm một phép biến đổi tối ưu dẫn tới một hàm khoảng cách tối ưu bằng cách cực đại hóa tổng phương sai giữa các phần tử khác tập mẫu (âm hoặc dương) và cực tiểu hóa phương sai của dữ liệu trong cùng tập mẫu (âm hoặc dương).

Giả sử tập ảnh kết quả khởi tạo gồm N ảnh: $X = \{x_i\}_{i=1}^N$ và một số các ràng buộc. Tập ảnh kết quả khởi tạo được trả về cho người dùng phản hồi liên quan và được chia thành hai tập phân biệt là tập mẫu dương (positive samples) và mẫu âm (negative samples). Để đạt được mục tiêu DCA, chúng ta cần xác định hai ma trận phương sai là $\hat{C}_b$ và $\hat{C}_w$ là khoảng cách giữa các kỳ vọng của các lớp khác nhau và khoảng cách giữa kỳ vọng và các mẫu của mỗi lớp. Được tính theo công thức sau:

$$\hat{C}_b = \frac{1}{n_b} \sum_{j=1}^{2} \sum_{i \in D_j} (m_j - m_i)(m_j - m_i)^T \quad (3)$$

$$\hat{C}_w = \frac{1}{n} \sum_{j=1}^{2} \frac{1}{n_j} \sum_{i=1}^{n_j} (x_{ji} - m_i)(x_{ji} - m_i)^T \quad (4)$$

Với $n_b$ là tổng số lượng phần tử của hai tập, $m_j$ là tâm của lớp j với $m_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ji}$, với $x_{ji}$ là véc tơ i của lớp j, mỗi $D_j$ là một lớp và trong bài toán này chúng ta có 2 lớp gồm tập mẫu dương và tập mẫu âm.

Ý tưởng của DCA là tìm một phép biến đổi tuyến tính mà đưa ra một hàm khoảng cách tối ưu bằng việc cực đại hóa tổng khoảng cách các kỳ vọng của các lớp khác nhau và cực tiểu hóa tổng khoảng cách các kỳ vọng trong cùng lớp. Quá trình DCA sẽ đưa về bài toán tối ưu như sau:

$$J(A) = argmax_A \frac{|A^T \hat{C}_b A|}{|A^T \hat{C}_w A|} \quad (5)$$

Ma trận A là ma trận biến đổi tối ưu mà chúng ta cần tìm. Khi tìm được phép biến đổi tối ưu A, chúng ta sẽ có được trọng số tối ưu của hàm khoảng cách Mahalanobis: $M = A^T A$.

Theo lý thuyết Fisher [11,12], bài toán tối ưu (5) tương ứng với việc cực đại hóa tổng khoảng cách các kỳ vọng của các lớp khác nhau và cực tiểu hóa tổng khoảng cách các kỳ vọng trong cùng lớp, tương ứng là $\hat{C}_b$ và $\hat{C}_w$ [10]. Để tìm được lời giải cho bài toán (5), bài báo đề xuất thuật giải sau, thuật giải cũng được dùng để giải các nghiên cứu trước đây về LDA [22].

**Thuật toán 1.2. Discriminative Component Analysis**

**Input:**

- Tập ảnh $X = \{x_i\}_{i=1}^N$

- Tập các mẫu (liên quan, không liên quan) $D_j = \{x_1, x_2, \ldots, x_{n_j}\}$ , j =1,2.

**Output:**

- Ma trận biến đổi tối ưu A

- Ma trận tối ưu Mahalanobis $M_o$

**1.** Tính ma trận $\hat{C}_b$ và $\hat{C}_w$ theo công thức (1.1) và (1.2)

**2.** Chéo hóa ma trận $\hat{C}_b$ sử dụng eigen analysis:

2.1. Tìm U sao cho: $U^T\hat{C}_bU = \wedge_b$; $U^TU = I$, $\wedge_b$ là ma trận đường chéo các thành phần là trị riêng của U;

2.2. Tìm $\hat{U}$ là k thành phần của $U$ với các thành phần trị riêng khác 0

2.3. Tìm $D_b = \hat{U}^T\hat{C}_b\hat{U}$ là ma trận vuông cấp k là ma trận con của ma trận $\wedge_b$.

2.4. Tìm $Z = \hat{U}D_b^{-1/2}$ và $C_z = Z^T\hat{C}_wZ$;

**3.** Chéo hóa ma trận $\hat{C}_z$ sử dụng eigeinanalysis:

3.1. Tìm V sao cho: $V^T\hat{C}_zV = \wedge_w$; $V^TV = I$, $\wedge_w$ là ma trận đường chéo các thành phần là trị riêng của U;

3.2. Nếu cần giảm chiều, giả sử số chiều mong muốn là *r* thì $\hat{V}$ chính là r véc tơ cột của ma trận V mà mỗi cột là véc tơ riêng tương ứng với giá trị riêng nhỏ nhất. Tìm $D_w = \hat{V}^TC_z\hat{V}$ ; với $\hat{V} = V$ và $D_w = \wedge_w$.

**4.** Ta có: A = $Z\hat{V}D_w^{-1/2}$ và $M_o = A^TA$.

Thuật toán 1.2 thực hiện như sau, ma trận U là ma trận chéo hóa của ma trận phương sai $\hat{C}_b$ thể hiện sự tách biệt giữa hai tập mẫu liên quan và không liên quan. Sau khi bỏ đi các véc tơ với trị riêng bằng 0, chúng ta có ma trận vuông cấp k là $D_b$ là ma trận đường chéo với thành phần trên đường chéo là các trị riêng khác 0 của ma trận $\hat{U}$. Sau đó, thuật toán sẽ thực hiện việc tìm phép chiếu $Z = \hat{U}D_b^{-1/2}$, phép chiếu này làm cho các lớp khác nhau có sự phân biệt lớn nhất. Tiếp theo, chúng ta tính $C_z = Z^T\hat{C}_wZ$ và tìm ma trận V để chéo hóa của ma trận $C_z$. Nếu muốn giảm chiều, giả sử số chiều mong muốn là *r* thì $\hat{V}$ chính là r véc tơ cột của ma trận V mà mỗi cột là véc tơ riêng tương ứng với giá trị riêng nhỏ nhất. Từ đó, cho chúng ta được ma trận đường chéo $D_w = \hat{V}^TC_z\hat{V}$ . Cuối cùng, chúng ta có ma trận biến đổi tối ưu A và ma trận tối ưu Mahalanobis M: A = $Z\hat{V}D_w^{-1/2}$ và $M_o = A^TA$.

### 3.4. Thuật toán tra cứu

Thuật toán 1.3 dưới đây là mô tả thuật toán tra cứu ảnh hiệu quả sử dụng với học biểu diễn ảnh dựa vào học sâu và kết hợp với hàm khoảng cách cải tiến Mahalanobis ***IRDLoM (Image Retrieval using Deep learning and optimal distance metric)*.**

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

**Input**:

Tập các ảnh: ***DB***

Ảnh truy vấn khởi tạo: **Q**

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

**Output**:

Tập kết quả được tra cứu: ***R***

1. S ←RIL<DB,*M*>;

2. $S_q$ ←RIL<Q,*M*>;

3. $Result_{Initial}(Q)$←***Retrieval*** $_{Initial}$ $(S_q, S, N)$

4. R←$Result_{Initial}(Q)$;

5. **Repeat**

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

5.2. $A = \boldsymbol{DCA}(F_{feature}, F^+_{label}, F^-_{label})$; *Tìm phép biến đổi tối ưu A*

5.3. $M_o = A^TA$; *Trọng số tối ưu của hàm khoảng cách mahalanobis*

5.4. *R*←***Ranking*** $(S, M_o, N)$*; Phân hạng lại tập ảnh theo hàm khoảng cách Mahalanobis với bộ trọng số tối ưu*

**until** (User dừng phản hồi);

6. **Return** R;

Thuật toán tra cứu ảnh hiệu quả sử dụng với học biểu diễn ảnh và kết hợp với hàm khoảng cách cải tiến Mahalanobis trên thực hiện như sau:

Mỗi ảnh trong tập ảnh DB được học biểu diễn (bước 1) và được biểu diễn bởi một véc tơ đặc trưng trong không gian đặc trưng nhiều chiều. Khi người dùng đưa vào ảnh truy vấn khởi tạo Q, thuật toán cũng sử dụng cùng một thủ tục để biểu diễn ảnh truy vấn cùng một cách với cơ sở dữ liệu ảnh để biểu diễn thành véc tơ đặc trưng ảnh truy vấn $S_q$ (bước 2). Truy vấn khởi tạo được thực hiện ở bước 3 bởi $Result_{Initial}(Q)$←***Retrieval*** $_{Initial}(S_q, S, N)$, ở đây $S_q$ là biểu diễn của ảnh truy vấn, S là tập biểu diễn của tập ảnh cơ sở dữ liệu và N là số các ảnh được tra cứu trong tập S sau mỗi làn lặp. Kết quả thực hiện tra cứu với truy vấn khởi tạo $Result_{Initial}(Q)$ được gán cho R (bước 4).

Trên tập $Result_{Initial}(Q)$; trả về bởi truy vấn khởi tạo, người dùng sẽ thực hiện lựa chọn những ảnh phù hợp với mong muốn của họ thông qua hàm ***Feedback***$(R)$ để được tập đặc trưng $F_{feature}$ và tập nhãn $F_{Label} = \{F^+_{label}, F^-_{label}\}$ (bước 5.1). Sau đó, thông tin phải hồi gồm tập phản hồi liên quan và không liên quan được đưa vào học DCA (bước 5.2) để tìm ra phép chiếu A bằng cách giải bài toán tối ưu (5). Kết quả của ma trận chiếu này được đưa vào để xây dựng ma trận trọng số tối ưu để cải tiến trọng số M của hàm khoảng cách Mahalanobis (bước 5.3). Lúc này, chúng ta có được hàm khoảng cách Mahalanobis cải tiến:

$$similarity(F_i, F_j) = (F_i - F_j)^TM(F_i - F_j)$$

Quá trình tra cứu sẽ thực hiện phân hạng lại toàn bộ tập ảnh trong cơ sở dữ liệu ảnh bởi hàm ***Ranking*** $(S, M, N)$ và lấy ra N ảnh làm tập kết quả trả về cho người dùng (bước 5.4).

## IV. ĐÁNH GIÁ THỰC NGHIỆM

*A. Cơ sở dữ liệu ảnh*

Để chứng minh hiệu quả của phương pháp đề xuất, thực nghiệm tiến hành trên cơ sở dữ liệu ảnh COREL gồm 10.800 ảnh. Một số hình được chỉ trong dưới. Trong tập cơ sở dữ liệu ảnh COREL, mỗi thư mục gồm 100 ảnh tập tin cây nền gồm 80 khái niệm khác nhau như hoa, hoàng hôn, tàu hỏa, xe hơi, xe buýt, bầu trời, biển...Tất cả các ảnh trong tập ảnh này có tính chất là đều chứa đối tượng nổi bật.

*Hình 2.Các mẫu trong cơ sở dữ liệu ảnh được gán nhãn.*

Chúng tôi kết hợp một đặc trưng màu 102 chiều và một kết cấu 88 chiều để biểu diễn các ảnh. Đặc trưng màu được cấu tạo bởi mô men màu 6 chiều, lược đồ màu 32 chiều và tương quan màu 64 chiều. Mô men màu có 6 chiều là bởi vì trong mỗi kênh màu H, S và V của không gian màu HSV, chúng tôi trích rút hai mô men là color mean, color Standard Deviation. Cũng trong không gian màu HSV, lược đồ màu được tính toán sử dụng 8*2*2 bins. Tương quan màu được tạo ra bởi sử dụng 4 bin cho mỗi kênh (R, G và B) trong không gian RGB. Đặc trưng kết cấu tích hợp các đặc trưng Gabor và các đặc trưng wavelet. Đặc trưng Gabor gồm Mean-squared energy và meanAmplitude cho 4 scale và 6 hướng cho ảnh đa cấp xám. đặc trưng wavalet 40 chiều gồm hai mô men của wavelet là trung bình, độ lệch chuẩn. Tóm lại, các đặc trưng này được tổ hợp thành một véc tơ đặc trưng có 190 giá trị (tức 6+32+64+40+48=190). Sau đó, tất cả các thành phần đặc trưng được chuẩn hóa thành các phân bố chuẩn với trung bình không và độ lệch chuẩn một để biểu diễn các ảnh. Các khoảng cách Euclid của các đặc trưng 190 chiều giữa ảnh truy vấn và các ảnh cơ sở dữ liệu được tính toán mà không sử dụng biến đổi. Các kết quả tra cứu này được gọi là "Baseline" cho các so sánh.

Bên cạnh đó, như đã trình bày ở phần trước, hầu hết các hệ thống CBIR đều phụ thuộc chủ yếu vào cách biểu diễn đặc trưng hình ảnh. Tuy nhiên với một hệ thống CBIR thông thường chỉ quan tâm đến cách biểu diễn ảnh bằng cách trích rút các đặc trưng toàn cục hoặc cục bộ một cách thủ công dẫn đến hiệu năng của hệ thống nghèo nàn. Do đó, chúng tôi sử dụng kỹ thuật học sâu học biểu diễn ảnh sử dụng mạng học sâu CNN tạo ra các đặc trưng mức cao từ hình ảnh.

Trong phương pháp đề xuất, chúng tôi sử dụng một mô hình CNN, có tên AlexNet [26], đã được tiền huấn luyện trên một tập dữ liệu rất lớn trên tập ImageNet, sau đó sử dụng mô hình làm khởi tạo để trích rút đặc trưng mức cao, còn được gọi là học biểu diễn ảnh. Mạng AlexNet có cấu trúc tương đối đơn giản, bao gồm 5 lớp tích chập và 3 lớp kết nối đầy đủ với các lớp giữa là các lớp lấy mẫu và ReLU, được huấn luyện song song trên hai card đồ họa GPU. Để phù hợp với bài toán tra cứu ảnh, chúng tôi chọn lớp FC 8 để trích rút các véc tơ để cho ra véc tơ đặc trưng có số chiều là 1000.

Lý do chính chúng tôi chọn cách này là tương đối hiếm khi chúng ta có một bộ dữ liệu đủ lớn để huấn luyện toàn bộ CNN từ đầu; ngoài ra, huấn luyện một mô hình CNN từ đầu sẽ mất rất nhiều thời gian. Các CNN thông thường được dùng cho các bài toán mang nhiệm vụ phân loại hình ảnh trong đó một hình ảnh được lan truyền qua mạng và xác suất cuối cùng được lấy từ lớp cuối của mạng. Tuy nhiên, trong quá trình học biểu diễn, thay vì cho phép hình ảnh lan truyền qua toàn bộ mạng, chúng ta có thể dừng việc truyền ở một lớp tùy ý, chẳng hạn như lớp được kết nối đầy đủ cuối cùng và trích rút các giá trị từ mạng tại thời điểm này, sau đó sử dụng chúng như các vectơ đặc trưng.

*B. Thực hiện truy vấn và đánh giá*

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

Hiệu quả tra cứu được đánh giá trên cơ sở dữ liệu ảnh COREL gồm 10.800 ảnh, tất cả các ảnh trong cơ sở dữ liệu được sử dụng để thực hiện các truy vấn. Thực nghiệm thực hiện đánh giá độ chính xác của phương pháp đề xuất dựa trên độ chính xác trung bình của 10.800 ảnh truy vấn. Mỗi truy vấn thực hiện sẽ trả về 100 ảnh, lý do

chọn 100 ảnh là bởi vì người dùng thường chỉ xem xét 2 trang màn hình và mỗi trang màn hình chứa 50 ảnh để lựa chọn ảnh phản hồi.

Nhằm mục đích đánh giá, bài báo sử dụng độ chính xác trung bình để đánh giá hiệu quả và so sánh với các phương pháp khác. Độ chính xác trung bình là tỷ lệ của số ảnh liên quan trong danh sách trả về cho người dùng và được tính toán bởi trung bình tất cả các truy vấn. Độ chính xác trung bình là tiêu chí đánh giá chính dùng để đánh giá độ chính xác so với các phương pháp khác. Độ lệch chuẩn dùng để đo lường độ biến thiên của độ chính xác trung bình.

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

Trong thực nghiệm, phương pháp đề xuất được so sánh với năm phương pháp tra cứu ảnh sử dụng các độ đo khoảng cách khác nhau: (1) **Euclid**: thực hiện tra cứu ảnh dựa vào độ đo khoảng cách Euclid (2) **Euclid cải tiến**: thực hiện tra cứu ảnh dựa vào độ đo khoảng cách Euclid có cải tiến trọng số của từng chiều đặc trưng; (3) **RCA**: thực hiện tra cứu với độ đo khoảng cách RCA được cải tiến từ độ đo khoảng cách Mahalanobis [8]; (4) **MCML**: thực hiện tra cứu ảnh với độ đo khoảng cách MCML được cải tiến từ độ đo khoảng cách Mahalanobis mà bộ trong số là kết quả của việc biến đổi dữ liệu với các ràng buộc nhãn và (5) phương pháp đề xuất ***IRDLoM*** thực hiện tra cứu trên bộ đặc trưng học sâu kết hợp với hàm khoảng cách mahalanobis tối ưu.

*Bảng I. So sánh độ chính xác trung bình của 5 phương pháp tại các mức Top-50, Top-100 sau 1 lần lặp phản hồi.*

| **Average prec.** | **Euclid** | **Eclid cải tiến** | **RCA** | **MCML** | **IRDLoM** |
|---|---|---|---|---|---|
| **Top 50 prec.** | 18.87 % | 26.01 % | 62.32 % | 64.02% | **66.32 %** |
| **Top 100 prec.** | 19.01 % | 26.08 % | 63% | 64.05% | **66.89 %** |

Như được chỉ ra trên Bảng I, phương pháp của chúng tôi cho độ chính xác cao gơn hẳn các phương pháp còn

lại. Lý do của điều này là bộ đặc trưng sâu trong phương pháp đề xuất đã bao gồm tính ngữ nghĩa của của ảnh và hàm khoảng cách của phương pháp đề xuất đã phản ảnh được khoảng cách thích hợp của các ảnh có chung một chủ đề.

## V. KẾT LUẬN

Bài báo này trình bày phương pháp ***IRDLoM***, một kỹ thuật tra cứu ảnh hiệu quả cho cải tiến hiệu năng của các hệ thống tra cứu ảnh đa điểm. ***IRDLoM*** tận dụng tốt thông tin của người dùng thông qua tập mẫu phản hồi liên quan và không liên quan thực hiện học một phép chiếu tối ưu nhằm mục đích phân tách các ảnh không liên quan và các ảnh liên quan gần nhau hơn. Từ đó, tìm ra được ma trận trọng số tối ưu của hàm khoảng cách Mahalanobis và sử dụng hàm khoảng cách cải tiến này thực hiện phân hạng toàn bộ tập ảnh cơ sở dữ liệu và trả về tập ảnh kết quả cho người dùng. Thực hiện thực nghiệm ***IRDLoM*** vào một cơ sở dữ liệu gồm 10800 ảnh minh chứng rằng ***IRDLoM*** cung cấp độ chính xác cao hơn hẳn so với các phương pháp **Euclid**, phương pháp **Euclid cải tiến,** phương pháp **RCA** [8] và phương pháp **MCML** [4]**.**

## TÀI LIỆU THAM KHẢO


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

[2] Ruigang Fu, Biao Li, Yinghui Gao, Ping Wang, (2016). Content-Based Image Retrieval Based on CNN and SVM, 2nd IEEE International Conference on Computer and Communications, 638-642.

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

[4] A. Globerson and S. Roweis. Metric learning by collapsing classes. Advances in Neural Information Processing Systems, 18:451, 2006.

[5] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. Advances in Neural Information Processing Systems, 18:1473, 2006.

[6] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall. Learning distance functions using equivalence relations. In ICML, pages 11–18, 2003.

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

[8] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, Learning a Mahalanobis Metric from Equivalence Constraints, in Journal of Machine Learning Research (JMLR), 2005.

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

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

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

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

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

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

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

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

[17] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In NIPS, 2005.

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

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

[20] E. Xing, A. Ng, and M. Jordan. Distancemetric learning with application to clustering with side-information. In NIPS, 2002.

[21] Flickner, M., Sawhney, H., Niblack, W., et al., (1995). Query by image and video content: The QBIC system. IEEE Computer Magazine 28 (9), 23–32.

[22] A. Pentland, R. W. Picard, and S. Sclaroff (1996). Photobook: content-based manipulation for image databases.International Journal of Computer Vision, 18(3):233–254.

[23] M. Ortega-Binderberger and S. Mehrotra (2004). Relevance feedback techniques in the MARS image retrieval systems. Multimedia Systems, 9(6):535–547.

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

[25] C. Carson, S. Belongie, H. Greenspan, and J. Malik (2002). Blobworld: image segmentation using expectation-maximization and its application to image querying. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(8):1026–1038, 2002.

[26] Krizhevsky, A., Sutskever, I., & Geoffrey E., H. (2012). ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25 (NIPS2012), 1–9.

[26] J. Z. Wang, J. Li, and G. Wiederhold, ( 2001). "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 23, no. 9, pp. 947-963.

# IMPROVE THE EFFECTIVENESS OF CONTENT-BASED IMAGE RETRIEVAL BY COMBINING DEEP LEARNING AND THE OPTIMAL DISTANCE FUNCTION

***Abstract:*** *Effective image representation and similarity measurement between two images are two important issues in improving the performance of a content-based image retrieval system. Deep learning has attracted the attention of researchers in the issue of effective image representation. Meanwhile, the problem of measuring the effective similarity towards learning distance measurement has an advantage. In this paper, we propose an image retrieval method, called IRDLoM (Image Retrieval using Deep learning and optimal distance metric). Method of representing images by deep features and measuring the similarity between two images by learning a measure of distance. The experimental results on the Corel photobook have proved the accuracy of the proposed method.*

**Đào Thị Thúy Quỳnh** nhận học vị tiến sĩ Máy tính, chuyên ngành Khoa học máy tính tại Học viện Khoa học và Công nghệ, Viện hàn lâm Khoa học và Công nghệ Việt Nam. Hiện nay, là giảng viên Khoa Công nghệ thông tin 1, Học viện Công nghệ Bưu chính Viễn thông.

Lĩnh vực nghiên cứu: Trí tuệ nhân tạo, học máy, xử lý ảnh, tra cứu ảnh dựa vào nội dung.

Email: quynhdao.ptit@gmail.com