MỘT PHƯƠNG PHÁP GOM CỤM DỮ LIỆU CHO BÀI TOÁN TÌM KIẾM ẢNH

Nguyễn Thị Thuỳ Trang, Trần Như Ý
Huỳnh Thị Châu Lan, Phan Thị Ngọc Mai*
Trường Đại học Công nghiệp Thực phẩm TP.HCM
*Email: *maiptn@hufi.edu.vn*
Ngày nhận bài: 05/3/2021; Ngày chấp nhận đăng: 24/5/2021

TÓM TẮT

Trong bài báo này, một cải tiến thuật toán K-Means được đề xuất nhằm phân cụm dữ liệu và áp dụng cho bài toán tìm kiếm ảnh tương tự theo nội dung. Để thực hiện được điều này, chúng tôi sử dụng một giá trị ngưỡng đo độ tương tự giữa các đối tượng dữ liệu, ngưỡng này được kí hiệu là θ. Trên cơ sở ngưỡng θ, thuật toán K-Means được cải tiến bằng cách không xác định trước số tâm cụm, số cụm dữ liệu tăng trưởng theo sự gia tăng của số lượng hình ảnh. Đặc trưng của hình ảnh được trích xuất dưới dạng một véc-tơ có *n* chiều và là dữ liệu đầu vào cho thuật toán K-Means đã được cải tiến để từ đó tìm kiếm các hình ảnh tương tự. Nhằm minh chứng cho các đề xuất, chúng tôi thực nghiệm và đánh giá kết quả trên tập dữ liệu ảnh COREL (có 1000 ảnh) đồng thời so sánh với các công trình khác đã được công bố gần đây trên cùng bộ dữ liệu. Theo như kết quả thực nghiệm, những đề xuất của chúng tôi là khả thi và có thể ứng dụng cho các hệ thống tìm kiếm ảnh khác nhau.

Từ khoá: Phân cụm, K-Means, độ đo tương tự, ảnh tương tự.

1. GIỚI THIỆU

Trong những năm gần đây, nhiều hệ thống tìm kiếm đã được công bố nhằm giải quyết bài toán tìm kiếm ảnh tương tự trong các cơ sở dữ liệu đa phương tiện. Có nhiều lĩnh vực khác nhau áp dụng kỹ thuật tìm kiếm ảnh như y tế, thời trang, hệ thống giám sát đối tượng, hệ thống thông tin địa lý, thư viện số... [1, 2], nhiều hệ thống tra cứu ảnh dựa trên nội dung CBIR (*Content-Based Image Retrieval*) đã được giới thiệu [3, 4].

Một số công trình tìm kiếm ảnh đã được công bố như: tìm kiếm ảnh dựa trên thuật toán K-Means [5], tìm kiếm ảnh dựa trên hình dạng, màu sắc, cấu trúc, đối tượng đặc trưng [6, 7]. Các công trình đã khảo sát tập trung vào kỹ thuật trích xuất đặc trưng, kỹ thuật đối sánh và tìm kiếm dựa trên các đặc trưng… nên các phương pháp này rất tốn kém nhiều chi phí về thời gian và bộ nhớ để đối sánh hai đối tượng hình ảnh, cần có một phương pháp tra cứu hình ảnh tương tự dựa trên một dữ liệu trung gian để từ đó truy hồi hình ảnh. Mặt khác, việc tìm kiếm dữ liệu trung gian cần sử dụng các phương pháp khai phá dữ liệu để tìm ra tập dữ liệu đại diện cho hình ảnh. Trong đó, phương pháp phân cụm là một trong những kỹ thuật quan trọng trong khai thác dữ liệu và đã được ứng dụng trong nhiều hệ thống tìm kiếm ảnh đã được phát triển cho các cơ sở dữ liệu lớn, một thuật toán tiêu biểu trong phân cụm phân hoạch là K-Means [8].

Tuy nhiên, kết quả phân cụm thu được từ thuật toán K-Means phụ thuộc nhiều vào việc khởi tạo số lượng cụm ban đầu, điều này ảnh hưởng đến độ chính xác của quá trình phân cụm, nghĩa là phụ thuộc vào số lượng tâm cụm đã được chọn ban đầu. Ngoài ra, nếu

bổ sung phần tử mới vào cụm thì thuật toán K-Means cần phải được xác định lại tâm cụm mới, điều này làm cho tốn nhiều chi phí trong quá trình thực thi. Bên cạnh đó, nếu dữ liệu tăng trưởng ngày càng lớn thì việc xác định trước số lượng tâm cụm ban đầu là không phù hợp bởi vì có thể dẫn đến hai phần tử trong cùng một cụm có khoảng cách khá lớn [9]. Do đó, trong bài báo này, nhóm tác giả đã cải tiến thuật toán K-Means và áp dụng cho bài toán tìm kiếm ảnh tương tự theo nội dung. Trong cải tiến này, số lượng tâm cụm không cần phải xác định trước mà tăng dần khi thỏa một điều kiện cho trước và theo sự tăng trưởng của bộ dữ liệu.

Trong bài báo này, chúng tôi sử dụng một giá trị ngưỡng θ để đánh giá độ tương tự giữa các đối tượng dữ liệu. Thuật toán K-Means được cải tiến bằng cách không cần xác định trước số tâm cụm và số cụm dữ liệu tăng trưởng theo sự gia tăng của số lượng hình ảnh. Trên cơ sở này, chúng tôi áp dụng thuật toán K-Means cải tiến để phân cụm các dữ liệu của hình ảnh để thực hiện bài toán tìm kiếm ảnh tương tự theo nội dung. Ngoài ra, chúng tôi đề xuất mô hình thực nghiệm và xây dựng ứng dụng thực nghiệm trên bộ ảnh COREL để đánh giá độ chính xác và tính khả thi cho những đề xuất.

Đóng góp của bài báo bao gồm: (1) Xây dựng một phương pháp gom cụm cải tiến dựa trên K-Means nhằm tạo ra một mô hình phân loại dữ liệu cũng như giúp quá trình tìm kiếm được hiệu quả về tốc độ và độ chính xác; (2) Xây dựng mô hình tìm kiếm ảnh tương tự theo nội dung dựa trên thuật toán K-Means cải tiến; (3) Xây dựng mô hình thực nghiệm và thực thi trên một bộ dữ liệu phổ dụng nhằm minh chứng tính đúng đắn của lý thuyết đề xuất.

2. CÁC CÔNG TRÌNH LIÊN QUAN

Nhiều công trình sử dụng phương pháp gom cụm dựa trên K-Means nhằm thực hiện bài toán tìm kiếm ảnh đã được công bố gần đây như: Sử dụng thuật toán K-Means kết hợp phân lớp SVM (*Support Vector Machine*) nhằm thực hiện tìm kiếm ảnh dựa trên các đặc trưng cấp thấp [10], gom cụm các véc-tơ đặc trưng dựa trên histogram và đánh chỉ mục ảnh để thực hiện tìm kiếm nhanh tập ảnh tương tự [11], phân đoạn ảnh và trích xuất đặc trưng dựa trên vùng sử dụng giá trị kỳ vọng và độ lệch chuẩn của không gian màu RGB nhằm phân cụm và tra cứu hình ảnh tương tự [12], phân cụm các véc-tơ đặc trưng dựa trên K-Medoids và tìm kiếm tuyến tính cục bộ dựa trên ứng viên lân cận gần nhất [13], sử dụng kỹ thuật phân cụm K-Means để truy xuất hình ảnh dựa trên nội dung [14], phân cụm dựa trên đặc trưng của hình ảnh bằng bảng băm kết hợp với phương pháp giảm số chiều PCA [15], lập chỉ mục ảnh và phân cụm bằng K-Means dựa trên đặc trưng cấp thấp trong không gian RGB với khoảng cách Euclid [16] …

Theo H.K.Maur và cộng sự (2019) đã tiếp cận thuật toán K-Means nhằm phân cụm các hình ảnh dựa trên đặc trưng cấp thấp gồm: màu sắc, hình dạng và cấu trúc. Trên cơ sở phân cụm, nhóm tác giả áp dụng phương pháp phân lớp SVM để phân loại từng nhóm đối tượng hình ảnh, từ đó thực hiện tra cứu ảnh. Theo thực nghiệm, phương pháp đề xuất của tác giả đã giải quyết hiệu quả cho bài toán tra cứu ảnh trên nhiều bộ dữ liệu khác nhau. Tuy nhiên, việc phân cụm của nhóm tác giả còn phụ thuộc vào số tâm cụm ban đầu và nếu như dữ liệu tăng trưởng thì việc phân cụm mất nhiều chi phí thời gian. Do đó, cần phải có một phương pháp phù hợp cho bộ dữ liệu tăng trưởng để không tái cấu trúc các cụm nhằm giảm chi phí thời gian [10].

Theo Juli Rejito và cộng sự (2017) đã thực hiện một hệ truy vấn ảnh dựa trên thuật toán K-Means bằng cách phân cụm các đặc trưng dưới dạng histogram; sau đó, nhóm tác giả thực hiện tạo định danh cho tập hình ảnh nhằm tra cứu nhanh tập ảnh kết quả. Kết quả thực nghiệm của bài báo đã đạt được độ chính xác 0,68, điều đó cho thấy phương pháp áp dụng của nhóm tác giả là hiệu quả và khả thi cho việc áp dụng trong các hệ thống tìm kiếm ảnh.

Tuy nhiên, việc phân cụm trên các đặc trưng dựa vào K-Means sẽ tốn thời gian xác định tâm cụm, cũng như số cụm ban đầu. Đồng thời, khi dữ liệu tăng trưởng sẽ tốn chi phí phân cụm lại và định danh các hình ảnh theo từng cụm [11].

Mohamed Ouhda và cộng sự (2018) sử dụng kỹ thuật K-Means để phân cụm các đặc trưng dựa trên vùng của hình ảnh. Đặc trưng này được trích xuất trên từng vùng đối tượng theo vị trí và màu sắc trong không gian RGB để từ đó thực hiện bài toán tra cứu ảnh theo nội dung. Theo như kết quả của nhóm tác giả, phương pháp đề xuất thực nghiệm hiệu quả trên nhiều bộ ảnh khác nhau và có độ chính xác cao. Tuy nhiên, phương pháp này tốn chi phí trong việc phân cụm [12].

Wei Zhang và cộng sự (2016) trình bày mô hình K-Medoids để thay thế K-Means trong việc phân cụm dữ liệu. Sau đó, một phương pháp tìm kiếm tuyến tính cục bộ dựa trên ứng viên lân cận gần nhất được thực hiện cho một truy vấn. Kết quả phương pháp này là một láng giềng bằng cách xếp hạng các lân cận với khoảng cách Euclide của các véc-tơ đặc trưng ban đầu. Kết quả thử nghiệm trên bộ ảnh CIFAR-10 cho thấy phương pháp được đề xuất là hiệu quả và cải thiện đáng kể độ chính xác của việc tìm kiếm. Tuy nhiên, việc phân cụm trên K-Medoids yêu cầu đưa vào số lượng cụm k sẽ tốn kém thời gian xác định tâm cụm khi dữ liệu ảnh tăng tưởng [13].

Mostafa G. Saeed và cộng sự (2017) đề xuất một phương pháp xây dựng véc-tơ đặc trưng để mô tả một hình ảnh. Đặc trưng này gồm 140 thành phần được lấy dựa trên histogram màu, moment màu, bộ lọc Gabor và các đặc trưng cấp thấp về cấu trúc bề mặt ảnh… Trên cơ sở thuật toán K-Means, nhóm tác giả thực hiện phân cụm hình ảnh dựa vào khoảng cách Euclide. Thực nghiệm được thực thi trên bộ dữ liệu IMPLIcity có 1000 hình ảnh màu và đánh giá trên 5 hình ảnh ngẫu nhiên. Tuy nhiên, việc thử nghiệm này chỉ kiểm tra trên bộ dữ liệu nhỏ nên không đánh giá được hiệu quả trong các bộ dữ liệu lớn. Ngoài ra, việc phân cụm trên các véc-tơ đặc trưng dựa vào K-Means sẽ tốn chi phí khi xác định lại tâm cụm khi dữ liệu tăng trưởng [14].

Tongtong Yuan và cộng sự (2019) đề xuất một phương pháp băm thích ứng nhằm giảm số chiều cho véc-tơ đặc trưng của hình ảnh để thực hiện tìm kiếm ảnh tương tự. Trong cách tiếp cận này, véc-tơ đặc trưng được loại bỏ các giá trị dư thừa nhưng vẫn đảm bảo đặc tính của đối tượng hình ảnh để từ đó giảm chi phí trong quá trình tìm kiếm ảnh. Tuy nhiên, phương pháp này tốn chi phí trong việc phân cụm vì phải xác định lại tâm [15]

Bằng cách sử dụng thuật toán phân cụm K-Means, Annrose và cộng sự (2016) đã đề xuất phương pháp lập chỉ mục ảnh dựa trên đa đặc trưng để thực hiện bài toán tìm kiếm ảnh. Nhóm tác giả thực hiện phương pháp K-NN kết hợp K-Means để tìm kiếm và đối sánh đặc trưng hình ảnh. Tuy nhiên, việc phân cụm của nhóm tác giả còn phụ thuộc vào số tâm cụm ban đầu và nếu như dữ liệu tăng trưởng thì phải xác định lại tâm cụm [16].

Nhằm tăng tốc độ tìm kiếm cho bài toán tìm kiếm hình ảnh tương tự, A. K. Jain đã khảo sát và mô tả phương pháp gom cụm là nhóm các đối tượng theo độ đo tương tự cho trước [17].

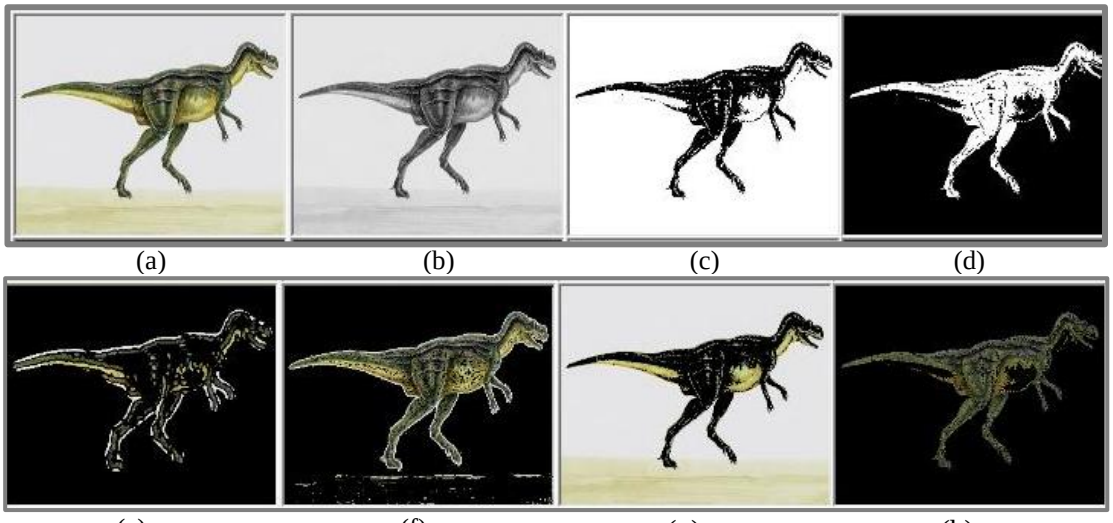
Phương pháp gom cụm đã được ứng dụng nhiều trong bài toán tìm kiếm ảnh tương tự như: ứng dụng thuật toán K-Means và khoảng cách Euclide để gom cụm các đặc tính ảnh màu nhằm phục vụ cho bài toán tìm kiếm ảnh [18]. Các kỹ thuật đã khảo sát đều tập trung vào việc phân cụm với số cụm cho trước và từ đó kết hợp với một phương pháp khác để thực hiện bài toán tra cứu ảnh. Tuy nhiên, nếu chọn trước số cụm thì hai đối tượng có thể khác nhau nhưng có thể thuộc về một cụm, điều này dẫn đến nhiều sai số khi áp dụng thuật toán K-Means cho bài toán tìm kiếm ảnh. Do đó, thuật toán K-Means cần phải được cải tiến để có thể tăng trưởng số cụm dữ liệu đồng thời đảm bảo 2 phần tử giống nhau phải thuộc về một cụm với một bán kính θ cho trước. Để thực hiện được vấn đề này, chúng tôi sử dụng một giá trị ngưỡng θ giữa các đối tượng dữ liệu và xây dựng một quy tắc phân bố các phần tử trên cơ

sở thuật toán K-Means. Từ đó, một thuật toán cải tiến được đề xuất sao cho không xác định trước số tâm cụm, số cụm dữ liệu tăng trưởng theo sự gia tăng của số lượng hình ảnh nhằm áp dụng cho bài toán tìm kiếm ảnh.

3. KỸ THUẬT PHÂN CỤM

3.1. Đặc trưng hình ảnh

Trong bài báo này, chúng tôi trích xuất đặc trưng dựa trên màu sắc, đối tượng và vị trí tương đối của đối tượng đặc trưng trên hình ảnh. Đặc trưng màu sắc được sử dụng dựa trên giá trị Histogram của 6 màu cơ bản gồm: đỏ (red), xanh lục (green), xanh dương (blue), vàng (yellow), cam (orange), tím (purple); Đặc trưng đối tượng và đặc trưng vị trí được trích xuất dựa trên độ tương phản bao gồm màu nền và màu đối tượng, đồng thời tính tỷ lệ về diện tích và chu vi của đối tượng. Trong Hình 1, các đặc trưng được trích xuất dựa trên màu sắc và vị trí tương đối của đối tượng đặc trưng.



Hình 1. Một ví dụ về trích xuất các vùng đặc trưng trên ảnh

Hình 1.a là ảnh gốc; Hình 1.b là ảnh lấy theo độ tương phản, nghĩa là nếu độ sáng của điểm ảnh dưới mức ngưỡng thì chuyển thành màu nền, ngược lại điểm ảnh đó lấy theo cường độ xám; Hình 1.c và Hình 1.d là ảnh mặt nạ của đối tượng và ảnh nền được lấy dựa trên độ tương phản; Hình 1.e là đường biên ảnh, được trích xuất theo phương pháp LoG (*Laplacian-of-Gaussian*); Hình 1.f là ảnh bề mặt được lấy theo phép lọc Sobel; Hình 1.g và Hình 1.h là ảnh đối tượng và ảnh nền.

Đặc trưng của một hình ảnh được trích xuất theo: tỷ lệ diện tích vùng, giá trị kỳ vọng theo trục X, giá trị kỳ vọng theo trục Y, độ lệch theo trục X, độ lệch theo trục Y, chu vi của đối tượng, màu sắc chính của ảnh gốc, màu sắc chính của đối tượng và hình nền. Trên cơ sở này, một vector đặc trưng có 44 chiều được trích xuất cho mỗi ảnh như sau:

Bảng 1. Một kết quả trích xuất đặc trưng cho Hình 1 của bộ COREL

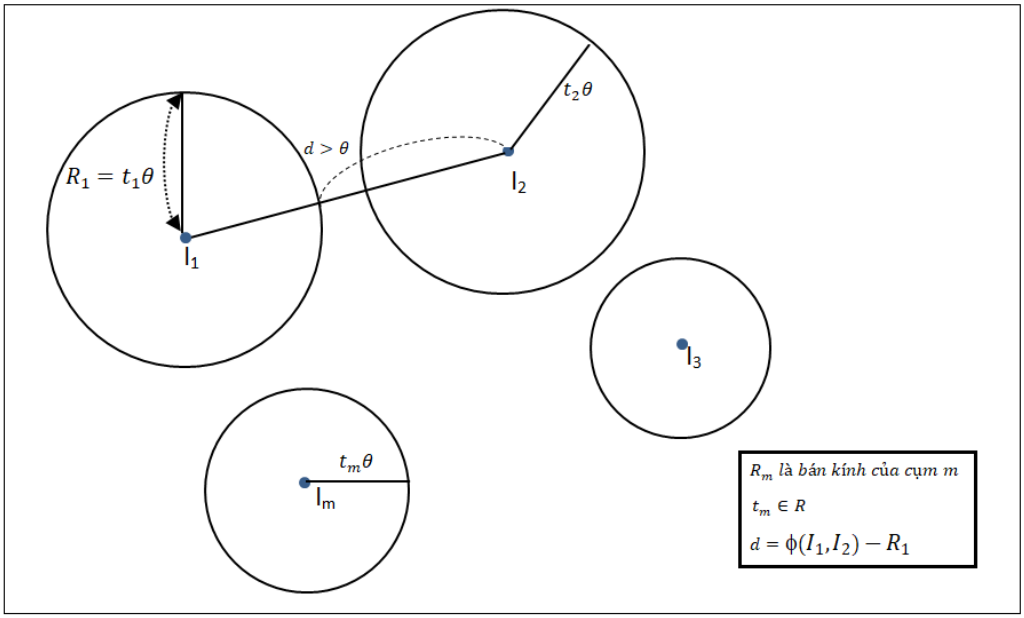
Tên đặc trưng	Giá trị
Diện tích hình đối tượng	0,483; 0,483; 0,422; 0,040; 0,043
Diện tích hình nền	0,516; 0,512; 0,568; 0,041; 0,050
Chu vi đối tượng	0,057; 0,519; 0,366; 0,041; 0,039
Bề mặt đối tượng	0,565; 0,498; 0,516; 0,0406; 0,047
Màu sắc của ảnh bề mặt đối tượng	0,001; 0,355; 0,067; 0,260; 0; 0,314
Đặc trưng màu sắc của ảnh đối tượng	0; 0; 0,337; 0,509; 0; 0,152
Đặc trưng màu sắc của hình nền	0,001; 0,579; 0; 0,015; 0; 0,403
Đặc trưng màu sắc của ảnh gốc	0,001; 0,299; 0,163; 0,254; 0; 0,282

3.2. Kỹ thuật phân cụm

3.2.1. Mô tả kỹ thuật phân cụm

Trong thuật toán K-Means, ba tham số cần được khởi tạo ban đầu gồm: số lượng cụm *k*, tâm cụm và độ đo tương tự; Ngoài ra, nếu bổ sung phần tử mới vào cụm thì phải xác định lại tâm cụm mới. Tuy nhiên, với một bộ dữ liệu bất kỳ cho trước, số lượng các cụm rất khó xác định cũng như việc tăng trưởng dữ liệu có thể làm gia tăng số lượng cụm, điều này gây ra tốn kém nhiều chi phí về thời gian và quá trình thực thi khi tái tạo lại số cụm.

Trong bài báo này, chúng tôi đề xuất một cải tiến thuật toán K-Means nhằm phân cụm dữ liệu và tăng trưởng số cụm theo bộ dữ liệu. Để thực hiện được điều này, chúng tôi sử dụng một giá trị ngưỡng đo độ tương tự giữa các đối tượng dữ liệu, ngưỡng này được kí hiệu là θ. Trên cơ sở ngưỡng θ, thuật toán K-Means được cải tiến bằng cách không xác định trước số tâm cụm, vì vậy số cụm dữ liệu tăng trưởng theo sự gia tăng của số lượng hình ảnh.



Hình 2. Mô tả kỹ thuật phân cụm

Trong Hình 2, phương pháp phân cụm dựa trên K-Means được mô tả, trong đó: I_i, R_i, t_i lần lượt là tâm, bán kính, hệ số dãn nở bán kính của cụm C_i.

Gọi $L = \{v_1, v_2, \dots, v_n\}$ là tập véc-tơ ảnh ban đầu

– **Bước 1:** Tạo cụm C_1 đầu tiên: $R_1 = \theta$ và $I_1 = v_1$

– **Bước 2:** Xét $v_i \in L$ với $i = 2, .., n$

Tìm cụm C_t thỏa $\phi(v_i, I_t) - R_t = min\{\phi(v_i, I_j) - R_j\}$ với $j = 1, .., k$ (trong đó k là số lượng cụm đã tạo)

$d = \phi(v_i, I_t) - R_t$

Nếu $(d \leq \theta)$ thì

$C_t = C_t \cup \{v_i\}$; //Thêm v_i vào cụm C_t

Nếu $\phi(v_i, I_t) > R_t$ thì

$R_t = \phi(v_i, I_t)$; // Cập nhật bán kính cụm C_t

Ngược lại //Tạo cụm mới C_k có tâm là v_i

$I_k = v_i$;

$R_k = \theta$;

$\Omega = \Omega \cup C_k$;

3.2.2. Thuật toán phân cụm

Theo phương pháp đề xuất như trên, thuật toán gom cụm ảnh được cải tiến từ K-Means được mô tả như sau:

Thuật toán CTIR

Đầu vào: Ngưỡng tương tự θ và L (tập véc-tơ ảnh ban đầu)

Đầu ra: Tập các cụm Ω

Function Clustering_theta(θ, L)

Begin

Khởi tạo $\Omega = \emptyset$;

Foreach $\langle v_i \rangle \in L$ **do**

If ($\Omega = \emptyset$) **then** //Tạo cụm đầu tiên C_1

$I_1 = v_1$;

$R_1 = \theta$;

$\Omega = \{C_1\}$;

Else

Tìm cụm $C_t \in \Omega$ thỏa $\phi(v_i, I_t) - R_t = min\{\phi(v_i, I_j) - R_j\}$

với $j = 1, .., k$ (trong đó k là số lượng cụm đã được tạo)

$d = \phi(v_i, I_t) - R_t$

If $(d \leq \theta)$ **then**

$C_t = C_t \cup \{v_i\}$; //Thêm v_i vào cụm C_t

If $\phi(v_i, I_t) > R_t$ **then**

$R_t = \phi(v_i, I_t)$; // Cập nhật bán kính cụm C_t

EndIf;

Else //Tạo cụm mới C_k có tâm là v_i

$I_k = v_i$;

$R_k = \theta$;

$\Omega = \Omega \cup C_k$;

EndIf;

EndIf;

EndFor;

Return Ω;

End.

Mệnh đề 1. Thuật toán CTIR có độ phức tạp $O(n * m)$, với n, m lần lượt là bộ dữ liệu và số cụm tạo ra.

Chứng minh: Giả sử có n phần tử trong bộ dữ liệu được phân thành m cụm. Thuật toán cần phải duyệt qua n phần tử dữ liệu, với mỗi phần tử cần phải duyệt qua m cụm để phân bố dữ liệu. Trong trường hợp xấu nhất, thuật toán CTIR có số lần duyệt là $n * m$ để phân bố các phần tử vào các cụm. Khi số phần tử dữ liệu lớn, giá trị n và m rất lớn thì độ phức tạp của thuật toán CTIR là $O(n * m)$

3.3. Thuật toán tìm kiếm ảnh

Trên cơ sở tập các cụm Ω đã được phân hoạch theo thuật toán **CTIR**, quá trình tìm kiếm ảnh được thực hiện bằng cách chọn ra cụm C_m có tâm gần nhất với ảnh tra cứu. Tuy nhiên, chúng tôi chọn thêm các cụm láng giềng của C_m dựa trên độ đo giữa các tâm cụm để tăng số lượng kết quả ảnh tra cứu. Khi đó các bước thuật toán tìm kiếm ảnh như sau:

- **Bước 1:** Tìm cụm C_m có tâm gần với véc-tơ ảnh tra cứu nhất.
- **Bước 2:** Tìm h cụm láng giềng với cụm C_m. Tập ¥ chứa h cụm láng giềng và cụm C_m.
- **Bước 3:** Tìm tập £ chứa tất cả các véc-tơ trong ¥.
- **Bước 4:** Sắp xếp £ tăng dần theo độ đo.

Thuật toán **SEIR**:

- **Đầu vào:** véc-tơ đặc trưng p (ảnh tìm kiếm), tập cụm Ω và ngưỡng tìm kiếm σ.
- **Đầu ra:** tập Ψ chứa các id (định danh) của các ảnh tương tự với ảnh tìm kiếm.

Function ClusterRetrieval(p , Ω , σ)

Begin

Khởi tạo Ψ = ∅;

Tìm cụm $C_k \in \Omega$: ϕ (p ,v_k) = min{ ϕ (p,v_i), i= 1, ... , m};

(với m là số lượng cụm, v_i là véc-tơ tâm của cụm C_i)

//Tìm h cụm láng giềng với C_k

Sắp xếp Ω tăng dần theo ϕ (v_t,v_k) - (C_t.R+C_k.R)

(với C_t, v_t là cụm và tâm của cụm thứ t, t=1, ... , m)

Khởi tạo £=∅;

If (ϕ (v_i,v_k) - (C_i.R + C_k.R)<σ) then

£= £ ∪C_i với i=0,..m-1.

EndIf

Sắp xếp tập £ theo ϕ (l ,p) với ∀l ∈ £

Tạo tập định danh hình ảnh Ψ theo thứ tự sắp xếp của tập £.

Return Ψ ;

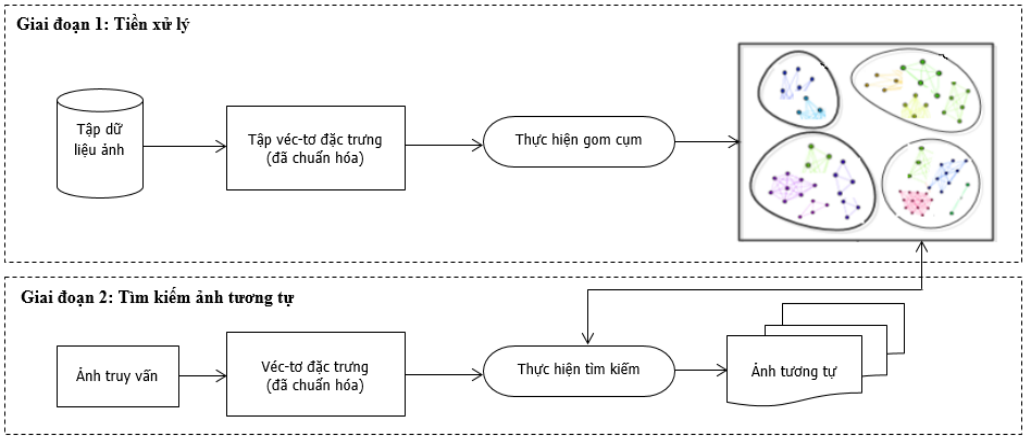
End.

Mệnh đề 2. Thuật toán SEIR có độ phức tạp $O(m\text{^}2)$, với m là số cụm tạo ra.

Chứng minh: Giả sử tập cụm đầu vào cho thuật toán SEIR có m cụm, khi đó thuật toán thực hiện việc tìm cụm gần nhất đối với ảnh đầu vào. Quá trình tìm kiếm này cần duyệt qua từng cụm, nghĩa là số lần so sánh là m tương ứng với m cụm. Sau khi thực hiện tìm cụm gần nhất với ảnh láng giềng, thuật toán SEIR tìm các cụm lân cận bằng cách sắp xếp lại tập cụm theo độ tương tự của cụm đã tìm được, số phép toán tối đa trong việc sắp xếp này là $O(m\text{^}2)$. Sau khi thực hiện tìm kiếm các cụm láng giềng, thuật toán thực hiện việc sắp xếp các hình ảnh theo độ đo tương tự với ảnh đầu vào (*tuy nhiên việc sắp xếp này có thể xử lý bên ngoài thuật toán nên độ phức tạp thuật toán SEIR có thể không bao gồm việc sắp xếp các hình ảnh theo độ đo tương tự*). Vì vậy, độ phức tạp của thuật toán SEIR là $O(m\text{^}2)$.

4. MÔ HÌNH TÌM KIẾM ẢNH

4.1. Mô tả mô hình



Hình 3. Mô hình của hệ thống tìm kiếm ảnh

Trong Hình 3 mô tả 2 giai đoạn được xử lý bao gồm: tiền xử lý để tạo dữ liệu cụm và tìm kiếm tập ảnh tương tự.

4.1.1. Tiền xử lý

– **Bước 1:** tạo véc-tơ đặc trưng thị giác cho mỗi hình ảnh trong tập dữ liệu ảnh.

– **Bước 2:** gom cụm các véc-tơ theo độ đo tương tự dựa trên thuật toán đã đề xuất.

4.1.2. Tìm kiếm ảnh tương tự

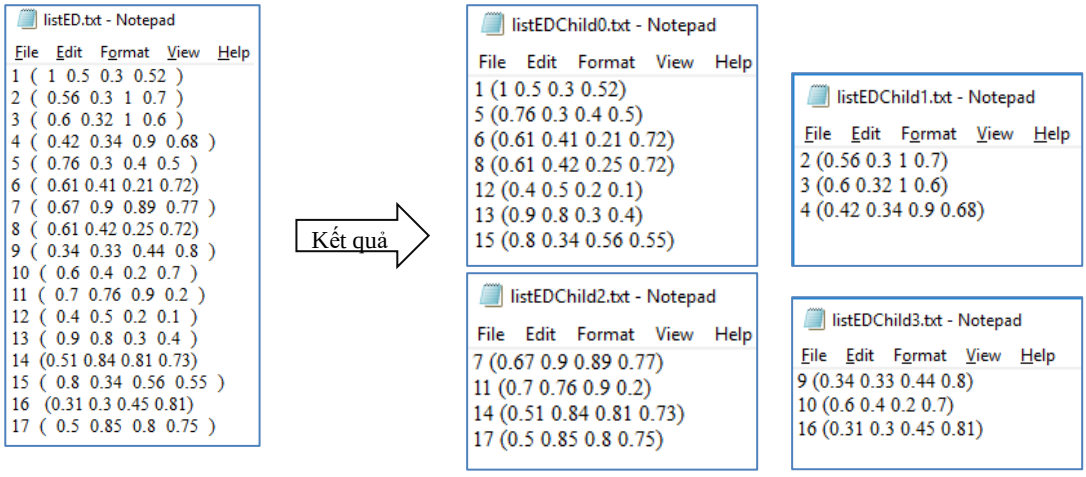
– **Bước 1:** từ một ảnh truy vấn, tạo véc-tơ đặc trưng cho ảnh này.

– **Bước 2:** thực hiện tìm kiếm một cụm gần nhất với ảnh truy vấn.

– **Bước 3:** kết xuất các ảnh kết quả sắp xếp theo độ đo tương tự với ảnh truy vấn.

4.2. Ví dụ thực nghiệm

Giả sử, bộ dữ liệu ảnh ban đầu gồm 17 ảnh biểu diễn bằng tập L gồm 17 véc-tơ đặc trưng. Chúng tôi tiến hành phân cụm bằng thuật toán cải tiến từ K-Means với ngưỡng $\theta = 0{,}2$. Kết quả sau khi phân cụm: số cụm thu được là 4 cụm như Hình 4.



Hình 4. Tập véc-tơ đặc trưng của 17 ảnh và kết quả phân cụm từ tập véc-tơ đặc trưng ban đầu

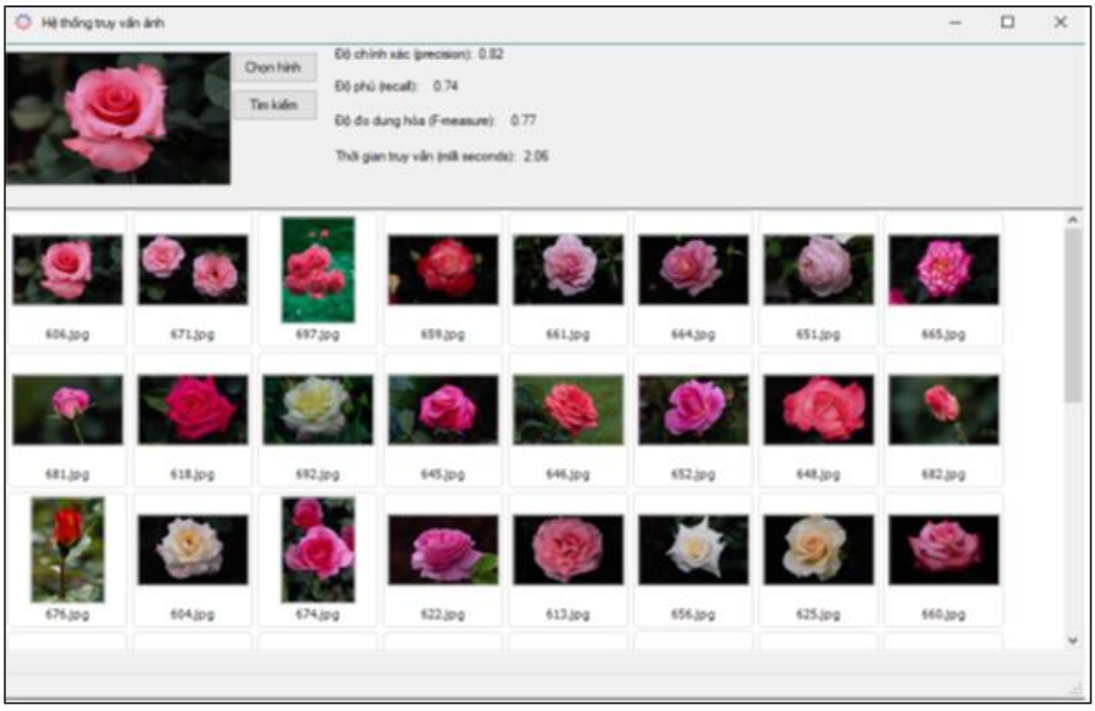
5. THỰC NGHIỆM

5.1. Môi trường thực nghiệm

Thực nghiệm gồm: (1) Giai đoạn tiền xử lý nhằm tạo ra tập véc-tơ đặc trưng cho tập dữ liệu hình ảnh; (2) gom cụm tập các véc-tơ dựa trên thuật toán đã được đề xuất; (3) tìm kiếm ảnh tương tự với một ảnh cho trước. Tất cả các ứng dụng thực nghiệm được xây dựng trên nền tảng dotNET Framework 4.5, ngôn ngữ lập trình C#. Các đồ thị được xây dựng trên Mathlab 2015. Cấu hình máy tính thực nghiệm: Core i3-7100U CPU @2.40GHz,8.0 GB RAM, hệ điều hành Windows 10 Pro 64 bit.

5.2. Ứng dụng thực nghiệm

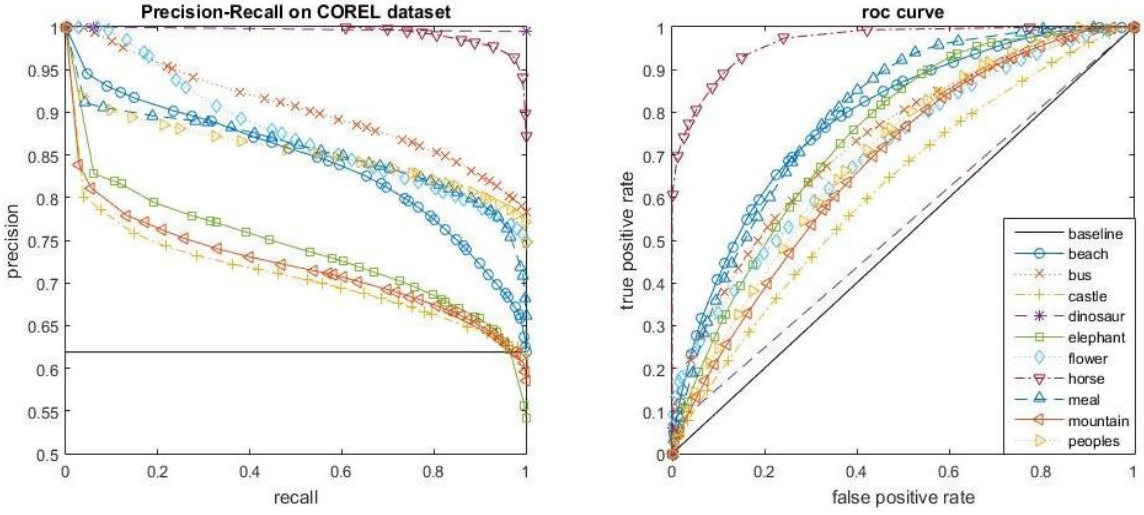
Trong bài báo này, chúng tôi tiến hành thực nghiệm trên bộ ảnh COREL có 1000 ảnh, kích thước 30.3 MB được chia thành 10 chủ đề: beach, bus, castle, dinosaur, elephant, flower, horse, meal, mountain, peoples. Hệ thống giúp truy vấn ảnh, với mỗi hình ảnh truy vấn sẽ được trích lọc trên tập dữ liệu ảnh COREL và tìm ra các hình ảnh có độ tương tự nhiều nhất với hình ảnh truy vấn như Hình 5.



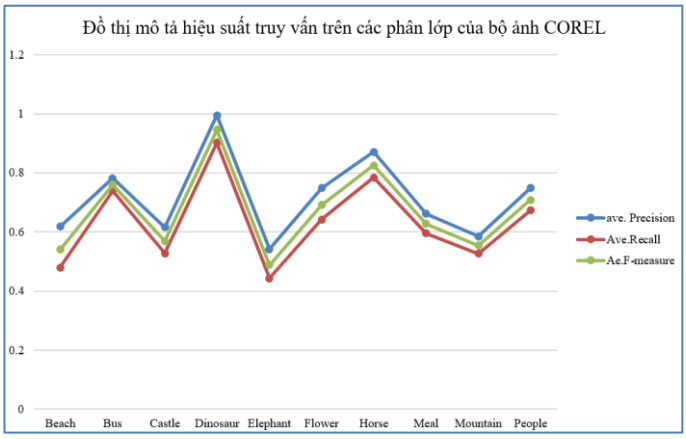
Hình 5. Một kết quả mẫu về truy vấn ảnh tương tự

5.3. Kết quả thực nghiệm

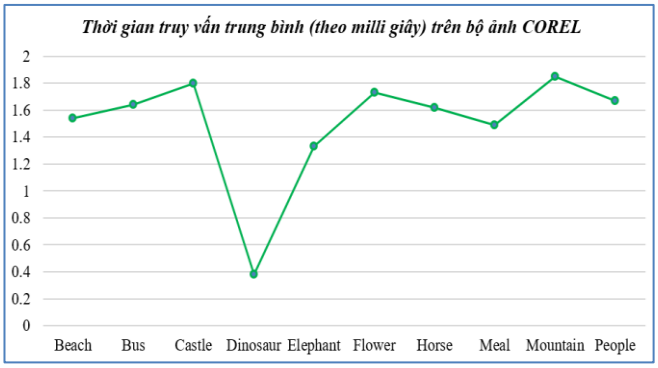
Mỗi đường cong trên đồ thị (hình 6) mô tả kết quả truy vấn độ chính xác (precision) và độ phủ (recall) tương ứng với từng tập dữ liệu ảnh theo phân lớp trong bộ dữ liệu COREL. Đồng thời, đường cong tương ứng trong đồ thị ROC cho biết tỷ lệ kết quả truy vấn đúng và sai, nghĩa là diện tích dưới đường cong này đánh giá được tính đúng đắn của các kết quả truy vấn.



Hình 6. Precision-Recall và đường cong ROC trên tập ảnh COREL



Hình 7. Giá trị trung bình của Precision, Recall, F-measure của tập dữ liệu COREL



Hình 8. Thời gian truy vấn trung bình (milliseconds) trên bộ ảnh COREL

Các giá trị về hiệu suất, thời gian tìm kiếm theo từng chủ đề và đánh giá so sánh cũng được trình bày cụ thể trong Bảng 2, 3 và 4. Theo như kết quả thực nghiệm trong các bảng này, phương pháp đề xuất của chúng tôi cho bài toán tìm kiếm ảnh tương tự là hiệu quả với độ chính xác trung bình là 71,70%.

Bảng 2. Bảng mô tả hiệu suất truy vấn trên các phân lớp của bộ ảnh COREL

Phân lớp ảnh	Ave. Precision	Ave.Recall	Ave.F-measure
Beach	0,618961457	0,48	0,540695031
Bus	0,782222222	0,7392	0,760102828
Castle	0,616233333	0,5282	0,568830769
Dinosaur	0,9944509	0,902225	0,946095707
Elephant	0,541933333	0,4434	0,48774
Flower	0,748883333	0,6419	0,691276923
Horse	0,871888889	0,7847	0,826
Meal	0,662222222	0,596	0,627368421
Mountain	0,584888889	0,5264	0,554105263
People	0,748444444	0,6736	0,709052632

Bảng 3. Bảng mô tả thời gian truy vấn trung bình trên bộ ảnh COREL

Phân lớp ảnh	Ave.time (milliseconds)
Beach	1,540088
Bus	1,640091
Castle	1,800112
Dinosaur	0,380025
Elephant	1,330072
Flower	1,7301
Horse	1,620093
Meal	1,490086
Mountain	1,8481034
People	1,670093

Bảng 4. Giá trị hiệu suất, thời gian truy vấn trung bình trên bộ ảnh COREL

Ave. Precision	Ave.Recall	Ave.F-measure	Ave.Time (milliseconds)
0,717012902	0,6315625	0,671126757	1,50488634

Kết quả thực nghiệm cho thấy, thuật toán CTIR thực hiện gom cụm theo ngưỡng θ đã xây dựng được chương trình tìm kiếm ảnh hiệu quả, nghĩa là thời gian tìm kiếm nhanh và có độ chính xác cao. Để minh chứng cho mô hình truy vấn ảnh được đề xuất là hiệu quả, chúng tôi so sánh kết quả thực nghiệm với một số công trình gần đây trên cùng bộ dữ liệu trong Bảng 5.

Bảng 5. So sánh hiệu suất truy vấn giữa các phương pháp trên bộ dữ liệu COREL

Phương pháp	Ave. Precision
A. Huneiti, 2015 [3]	55,88%
Bella M. I. T., 2019 [4]	60,90%
Phương pháp đề xuất	71,70%

Kết quả thực nghiệm cho thấy, phương pháp của chúng tôi đề xuất có độ chính xác trung bình là 71,7% và thời gian tìm kiếm trung bình là 1,5 milli giây. So sánh kết quả này với các phương pháp khác trên cùng một bộ dữ liệu mẫu thì thấy phương pháp tra cứu ảnh của chúng tôi đề xuất có độ chính xác cao hơn hai phương pháp: A. Huneiti (2015) với độ chính xác là 55,88% [3] và Bella M.I.T (2019) với độ chính xác là 60,90% [4]. Thuật toán K-Means được cải tiến bằng cách không xác định trước số tâm cụm, vì vậy khi tăng số lượng phần tử ảnh thì số cụm sẽ tăng trưởng theo thay vì phải gom cụm lại từ đầu như thuật toán K-Means, giúp giảm thời gian của quá trình gom cụm. Tuy nhiên, khi xuất hiện các cụm có quá nhiều phần tử sẽ ảnh hưởng đến độ chính xác trong quá trình truy vấn. Chẳng hạn, phân lớp Elephant có độ chính xác truy vấn tương đối thấp (48,77%).

6. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Bài báo đã đề xuất một cải tiến thuật toán K-Means thực hiện gom cụm nhằm tăng hiệu suất tìm kiếm ảnh tương tự. Trên cơ sở lý thuyết đã được đề xuất, nhóm tác giả xây dựng hệ

truy vấn ảnh theo nội dung. Kết quả thực nghiệm trên bộ dữ liệu ảnh COREL được đánh giá và so sánh với các công trình khác trên cùng một tập dữ liệu ảnh đã cho thấy phương pháp đề xuất là hiệu quả. Trên cơ sở ngưỡng θ, thuật toán K-Means được cải tiến bằng cách không xác định trước số tâm cụm. Vì vậy, số cụm dữ liệu tăng trưởng theo sự gia tăng của số lượng hình ảnh đã giảm được đáng kể thời gian của quá trình gom cụm so với thuật toán K-Means. Tuy nhiên, việc này dẫn đến xuất hiện các cụm có quá nhiều phần tử, ảnh hưởng đến độ chính xác của hệ thống. Hướng phát triển tiếp theo của nghiên cứu là xây dựng thuật tách cụm lớn thành 2 cụm nhỏ nhằm đảm bảo các phần tử trong cùng một cụm phải tương tự nhau.

Lời cảm ơn: Nhóm tác giả chân thành cảm ơn Trường Đại học Sư phạm TP. HCM, Trường Đại học Công nghiệp Thực phẩm TP. HCM là những nơi bảo trợ cho nghiên cứu này. Chúng tôi trân trọng cảm ơn nhóm nghiên cứu SBIR-HCM đã hỗ trợ về chuyên môn giúp chúng tôi hoàn thành bài nghiên cứu này.

TÀI LIỆU THAM KHẢO

1. Muneesawang P., Zhang N., Guan L. - Multimedia database retrieval: Technology and applications, Graduate Texts in Mathematics, Springer, New York Dordrecht London (2014).
2. Xie X., Cai X., Zhou J., Cao N., & Wu Y. - A semantic-based method for visualizing large image collections, IEEE Transactions on Visualization and Computer Graphics **25** (7) (2018) 2362-2377.
3. Huneiti A., Daoud M. - Content-based image retrieval using SOM and DWT, Journal of software Engineering and Applications **8** (02) (2015) 51.
4. Bella M. I. T., & Vasuki A. - An efficient image retrieval framework using fused information feature, Computers & Electrical Engineering **75** (2019) 46-60.
5. Lin C.-H., Chen C.-C., Lee H.-L., Liao J.-R. - Fast K-means algorithm based on a level histogram for image retrieval, Expert Systems with Applications **41** (7) (2014) 3276-3283.
6. Kim S., Park S., Kim M. - Central object extraction for object-based image retrieval, In: Bakker E.M., Lew M.S., Huang T.S., Sebe N., Zhou X.S. (eds) Image and Video Retrieval, CIVR 2003, Lecture Notes in Computer Science 2728, Springer (2003) 39-49.
7. Yoo H.W., Jung S.H., Jang D.S., Na Y.K. - Extraction of major object features using VQ clustering for content-based image retrieval, Pattern Recognition **35** (5) (2002) 1115-1126.
8. Kumar R.R., Prasad A.B. - K means clustering algorithm for partitioning data sets evaluated from horizontal aggregations, IOSR Journal of Computer Engineering **12** (5) (2013) 45-48.
9. Yadav A., Sing S.K. - An improved K-Means clustering algorithm, International Journal of Computing Academic Research **5** (2) (2016) 88-103.
10. Maur Harleen Kaur, Puneet Jain - Content based image retrieval system using K-Means clustering algorithm and SVM classifier technique, International Journal of Advance Research, Ideas and Innovations in Technology **5** (2) (2019) 39-43.
11. Juli Rejito, Atje Setiawan Abdullah, Akmal, Deni Setiana and Budi Nurani Ruchjana - Image indexing using color histogram and k-means clustering for optimization

CBIR in image database, The Asian Mathematical Conference 2016 (AMC 2016), IOP Conf. Series: Journal of Physics: Conf. Series 893 (2017) 012055.

12. Mohamed Ouhda, Khalid El Asnaoui, Mohammed Ouanan and Brahim Aksasse - A content based image retrieval method based on K-Means clustering technique, Journal of Electronic Commerce in Organizations **16** (1) (2018) 82-96.
13. Wei Zhang, Lihua Tian, Shanmin Pang, Chen Li - Multiple Cartesian K-medoids for a fine quantization, IEEE 22nd International Conference on Parallel and Distributed Systems (2016) 1216-1220.
14. Mostafa G. Saeed, Fahad Layth Malallah, Zaid Ahmed Aljawaryy - Content-based image retrieval by multifeatures extraction and K-Means clustering, International Journal of Electrical, Electronics and Computers (EEC Journal) **3** (2017) 1-11.
15. Tongtong Yuan, Weihong Deng, Jiani Hu, Zhanfu An, Yinan Tang - Unsupervised adaptive hashing based on feature clustering, Neurocomputing (2018) 1-41.
16. Annrose J., Christopher C.S. - Content based image retrieval using query based feature reduction with K-means cluster index, Asian Journal of Research in Social Sciences and Humanities **6** (12) (2016) 852-872.
17. Jain A.K. - Data clustering: 50 years beyond K-means, Pattern Recognition Letters **31** (8) (2010) 651-666.
18. Lin C.H., Chen C.C., Lee H.L., & Liao J.R. - Fast K-means algorithm based on a level histogram for image retrieval, Expert Systems with Applications **41** (7) (2014) 3276-3283.

ABSTRACT

A METHOD OF CLUSTERING FOR CONTENT-BASED IMAGE RETRIEVAL

Nguyen Thi Thuy Trang, Tran Nhu Y
Huynh Thi Chau Lan, Phan Thi Ngoc Mai*
Ho Chi Minh City University of Food Industry
*Email: *maiptn@hufi.edu.vn*

In this paper, an improvement in K-Means algorithm was proposed to cluster and applied to the problem of searching similar images by content. To accomplish this, we used a threshold value that measured the similarity between data objects, which is called as θ. K-Means algorithm was improved by not pre-determining the number of cluster centers, the number of data clusters grow with the increase in the number of images. The image was extracted as a n-dimensional vector and was an input for the improved K-Means algorithm from which to search for similar images. In order to demonstrate the proposals, we experimented and evaluated the results on the COREL image data set (1000 images) and compared to other recently published works on the same dataset. According to the experimental results, our proposals are feasible and applicable to different image retrieval systems.

Keywords: Cluster, K-Means, similarity measure, similar images, image retrieval.