MỘT CẢI TIẾN THUẬT TOÁN FCM ĐỂ PHÂN CỤM DỮ LIỆU LỚN VÀ ỨNG DỤNG CHO TRUY VẤN ẢNH

Hoàng Văn Quý[1*], Ngô Hoàng Huy[2], Nguyễn Thế Cường[1],
Nguyễn Tu Trung[3], Nguyễn Văn Quyền[4]

Tóm tắt: *Thuật toán đánh hạng đa tạp EMR được sử dụng rộng rãi trong truy vấn ảnh với cơ sở dữ liệu lớn. Việc xây dựng đồ thị quan hệ kề nhờ các điểm neo trên không gian vector các đặc trưng ảnh mức thấp đã được thiết kế hiệu quả dựa trên phân cụm K-means. Bài báo giới thiệu kết quả mới sử dụng phép phân cụm mờ C-Means (FCM) cải tiến thay thế cho K-means để chọn các điểm neo trong tập ảnh cần tra cứu. Thực nghiệm đã chứng tỏ tính hiệu quả của đề xuất cách chọn tập các điểm neo cho thuật toán EMR bằng phương pháp phân cụm mới đã thực sự tăng chất lượng truy vấn ảnh.*

Từ khóa: Big data; EMR; K-means; FCM; CBIR.

1. MỞ ĐẦU

Độ đo đánh hạng đa tạp [1-5] đo độ tương tự của các ảnh được sử dụng rộng rãi trong CBIR. Với một giả định rằng, mỗi điểm dữ liệu trong một không gian đặc trưng có một mối quan hệ với các điểm dữ liệu khác tương tự trong không gian, thuật toán trước hết xây dựng một đồ thị có trọng số cho tất cả các điểm dữ liệu trong không gian đặc trưng, ở đó, mỗi cạnh được gán một trọng số để biểu diễn mối liên quan dữ liệu giữa hai điểm. Đầu tiên, điểm dữ liệu truy vấn ban đầu được gán một giá trị nhất định, các điểm dữ liệu còn lại có liên quan được gán giá trị 0. Thứ hai, tất cả các điểm dữ liệu lan truyền xếp hạng của chúng đến các điểm dữ liệu bên cạnh thông qua các đồ thị có trọng số. Quá trình lan truyền của các điểm số xếp hạng lặp đi lặp lại cho đến khi hội tụ tới một tình trạng ổn định toàn cục. Các điểm chính thức được xếp hạng đại diện cho việc giống nhau giữa điểm dữ liệu và điểm truy vấn. Các điểm dữ liệu tương tự như các điểm truy vấn là những điểm xếp hạng lớn nhất. Để tăng hiệu quả tính toán EMR đã sử dụng các điểm neo thay thế cho việc xét toàn bộ tập dữ liệu ảnh. Các điểm neo được xác định bằng tâm của các cụm thu được sau khi sử dụng thuật toán phân cụm K-means trên tập dữ liệu gốc.

Trong bài báo này, chúng tôi đề xuất một phương pháp chuẩn xác định các điểm neo để tăng hiệu quả của thuật toán đánh hạng đa tạp EMR bằng thuật toán phân cụm mờ C-means (FCM) cải tiến. Thuật toán đề xuất có thể phân cụm được hiệu quả khi số cụm rất lớn (có thể lên đến 10% cho tới 20% của số phần từ của tập dữ liệu).

Phần còn lại của bài báo được tổ chức như sau. Phần 2, một số nghiên cứu liên. Phần 3 là đề xuất thuật toán xác định các điểm neo theo tiếp cận phân cụm sử dụng thuật toán FCM. Các kết quả thực nghiệm đưa ra trong phần 4. Kết luận và hướng nghiên cứu tiếp theo được trình bày trong phần 5.

2. NGHIÊN CỨU LIÊN QUAN

Thuật toán đánh hạng đa tạp EMR

Bước quan trọng của thuật toán EMR là xác định giá trị thứ hạng độ đo tương tự của ảnh trong CSDL với ảnh truy vấn là sử dụng độ đo đa tạp chỉ cho các vector ảnh neo thay vì trên toàn bộ cơ sở dữ liệu ảnh.

Trong thuật toán EMR quan hệ kề nhau của hai vector ảnh được xây dựng dựa trên các điểm neo (anchor) thay vì dựa trên quan hệ s-láng giềng của từng vector ảnh, nghĩa là E_i gọi là được nối với E_j nếu $i \neq j$ và tồn tại một điểm neo chung A_c nào đó sao cho A_c là s-láng giềng của mỗi E_i và E_j;

Với mỗi vector ảnh E_i ký hiệu:

$$Nb(i,s) = nbest(\|E_i, A_l\|; s),\ d_s = \max_{l \in Nb(i,s)} \{d(E_i, A_l)\} \quad (1)$$

Và $K(t) = \begin{cases} \frac{3}{4}(1-t^2), if\ |t| \le 1 \\ 0, otherwise \end{cases}$

Trong đó: *nbest(i)* là vector thứ i trong nbest vector gần A_c nhất; $d(E_i,A_l)$ là khoảng cách giữa 2 vector A,B có cùng kích thước.

Độ đo đa tạp được xây dựng nhờ giải hàm mục tiêu sau:

$$EMR(r;Q) = \frac{1}{2}\left(\sum_{1 \le i,j \le n+1} w_{ij} \left\| \frac{r_i}{\sqrt{D_{ii}}} - \frac{r_j}{\sqrt{D_{jj}}} \right\|^2 + \mu \sum_{i=1}^{n+1} \|r_i - r_{0,i}\|^2 \right) \to \min \quad (2)$$

$$Z = (z_{ki})_{1 \le k \le C, 1 \le i \le n+1},\ z_{ki} = \frac{K\left(\frac{\|E_i, A_k\|}{d_s}\right)}{\sum_{l \in NB(i,s)} K\left(\frac{\|E_i, A_l\|}{d_s}\right)} \forall k \in Nb(i,s), z_{ki} = 0 \forall k \notin Nb(i,s) \quad (3)$$

$$W = (w_{ij})_{1 \le i,j \le n+1},\ W \overset{def}{=} Z^T Z,\ w_{ij} = \sum_{k \in Nb(i,s) \cap Nb(j,s)} z_{ki} * z_{kj}, 1 \le i,j \le n+1 \quad (4)$$

$$D_{ii} \overset{def}{=} \sum_{j=1}^{n+1} w_{ij} \quad (5)$$

Trong đó: Q là tập ảnh truy vấn (để thuận tiện gán $E_{n+1} = Q$); $r_{0,n+1} = 1$; $r_{0,i} = 0$; $i = \overline{1,n}$.

Như trình bày ở trên, kết quả đánh hạng của thuật toán EMR phụ thuộc vào việc chọn số lượng điểm neo C và tập điểm neo $\{A_c\}_{c=1}^{C}$. Trong [1] các tác giả đã sử dụng thuật toán phân cụm K-means để chọn các điểm tâm cụm là điểm neo. Qua thực nghiệm, chúng tôi thấy khi số cụm C chọn không đủ lớn thì EMR cho kết quả sai lệch khá nhiều. Vì vậy, chúng tôi đề xuất thuật toán phân cụm dựa trên FCM thay thế cho thuật toán K-means để chọn các điểm neo.

3. KỸ THUẬT ĐỀ XUẤT

Xác định các điểm neo bằng thuật toán FCM

Cho trước một cơ sở dữ liệu (CSDL) đặc trưng mức thấp $E = \{E_i\}_{1 \le i \le n}$, sử dụng FCM (về FCM, xem [7,8,9,10,11]) ta phân cụm E thành C cụm, thuật toán lặp FCM cực tiểu hóa hàm mục tiêu sau:

$$J(A,\mu) = \sum_{i=1}^{n} \sum_{c=1}^{C} \mu_c^p \|E_i - A_c\|^2 \to \min \quad (6)$$

Trong đó, các hằng số p > 1, $C \in N^+, C \ge 2$, $m = \dim(E_i)$, $\forall 1 \le i \le n$, độ đo khoảng cách Euclid, $\|E_i - A_c\|^2 = \sum_{j=1}^{m} (E_{i,j} - A_{c,j})^2$ và các ràng buộc cho ma trận độ thuộc $\{\mu_{c,i}\}$ không âm được cho như sau:

$$\forall i = \overline{1,n}, \sum_{c=1}^{C} \mu_{c,i} = 1.0 \tag{7}$$

$$\forall c = \overline{1,\mathrm{C}}, \sum_{i=1}^{n} \mu_{c,i} > 0. \tag{8}$$

Công thức lặp giải hàm mục tiêu (4) được cho như sau:

$$\forall c = \overline{1,C}, i = \overline{1,n}, \mu_{c,i} = \frac{1}{\left\{ \sum_{i'=1}^{n} \frac{\|E_i - A_c\|}{\|E_{i'} - A_c\|} \right\}^{\frac{2}{p-1}}} \tag{9}$$

$$\forall c = \overline{1,C}, \mathrm{A}_c = \frac{\sum_{i=1}^{n} \mu_{c,i}^{p} E_i}{\sum_{i=1}^{n} \mu_{c,i}^{p}} \tag{10}$$

Trong công thức (9) và (10) ở trên, tất cả các vector đặc trưng của CSDL ảnh đều tham gia. Nhưng thực tế là không phải tất cả các phần tử điểm ảnh đều có ảnh hưởng đến quá trình hiệu chỉnh tâm cụm, như trong các thuật toán K-means[6] thì chỉ có các phần tử gần A_c nhất mới tham gia vào việc hiệu chỉnh tâm.

Do các đặc điểm trên, chúng tôi đề xuất cải tiến các công thức (9) và (10) như sau:

$$\forall c = \overline{1,C}, i = \overline{1,n}, \mu_{c,i} = \min \left\{ \frac{1}{\left\{ \sum_{i'=1}^{n} \frac{\|E_i - A_c\|}{\|E_{i'} - A_c\|} \right\}^{\frac{2}{p-1}}}, \mu_{\varepsilon} \right\} \tag{11}$$

$$\forall c = \overline{1,C}, \mathrm{A}_c = \frac{\sum_{i'=1}^{nb} \mu_{c,\mathrm{nbest(i')}}^{p} E_{\mathrm{nbest(i')}}}{\sum_{i'=1}^{nb} \mu_{c,\mathrm{nbest(i')}}^{p}} \tag{12}$$

Trong đó: μ_ε là một hằng số dương đủ nhỏ; nb là tham số chỉ số các vector đặc trưng của CSDL ảnh E gần V_c nhất; nbest(i) là vector thứ i trong nbest vector gần A_c nhất.

Chúng ta có thuật toán FCM cải tiến để xác định các điểm neo như sau:

Thuật toán. (Xác định các điểm neo bằng FCM)-AFCM

Input: E=$\{E_i\}_{1\le i\le n}$ dữ liệu đặc trưng mức thấp, hằng số p > 1, C số điểm neo (C có thể rất lớn từ 10% đến 20% của số các ảnh của E), $m = \dim(E_i), i = \overline{1,n}$, nb số nbest các điểm gần một điểm neo A_c, $c = \overline{1,C}$, μ_ε> 0 và L là số vòng lặp tối đa.

Output: Tập các điểm neo $\{A_c\}_{1\le c\le C}$.

Bước 1: Khởi tạo các tâm cụm bằng thuật toán K-means:

1.1: Gọi thuật toán phân cụm K-mean để phân cụm E thành C cụm, thu được $\{A_c\}_{1\le c\le C}$.

1.2: Sử dụng thuật toán FCM gốc để giải hàm mục tiêu theo công thức (6).

Bước 2: Lặp $l = \overline{1, L}$:

2.1: Tính ma trận độ thuộc $\{\mu_{c,i}\}_{1 \le c \le C, 1 \le i \le n}$ theo công thức (11).

2.2: Chuẩn hóa các trọng số $\{\mu_{c,i}\}_{1 \le c \le C, 1 \le i \le n}$ theo ràng buộc (7): =>

$$\mu_{c,i} = \frac{\mu_{c,i}}{\sum_{c'=1}^{C} \mu_{c',i}}, c = \overline{1, C}, i = \overline{1, n}$$

2.3: Tính lại tâm các cụm $\{A_c\}_{1 \le c \le C}$ theo công thức (12).

2.4: Tính $J_l(A, \mu)$ theo công thức (6).

2.5: Ra khỏi vòng lặp nếu $J_l(A, \mu) \approx J_{l-1}(A, \mu)$ (sai số dưới ngưỡng).

Bước 3: Trả về $\{A_c\}_{1 \le c \le C}$.

Độ phức tạp của thuật toán là: O(n*m*C*L + n*m*C^2*L*N_best).

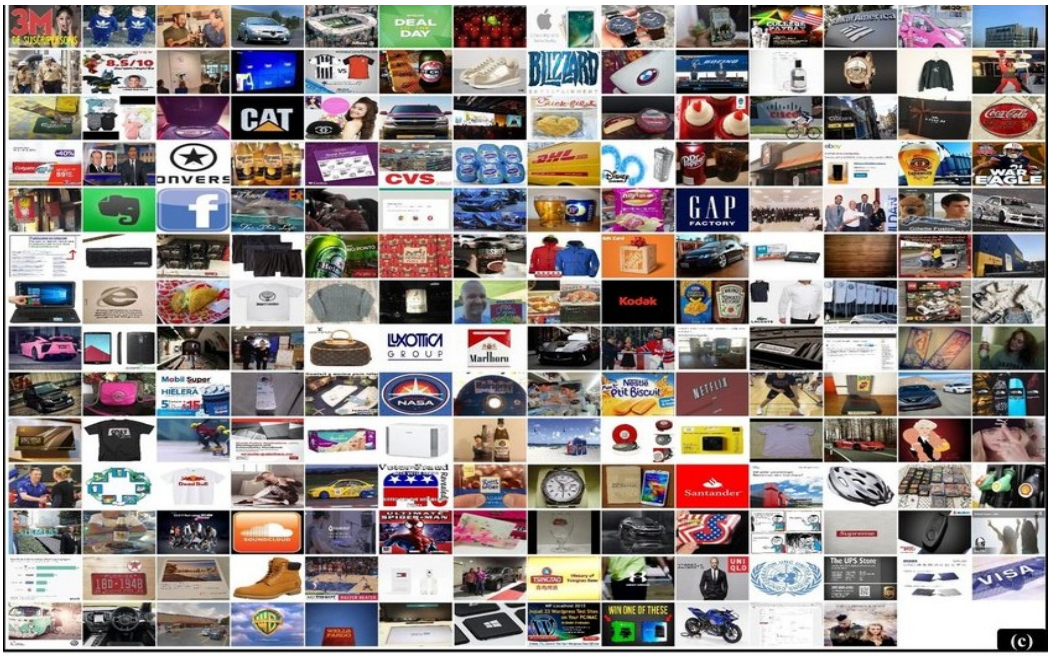
4. THỰC NGHIỆM

Chúng tôi tiến hành các thực nghiệm trên hai tập dữ liệulogo-2K+[12]vàVGGFACE2-S[13]. Các tập dữ liệu này được tổ chức thành các lớp ngữ nghĩa theo cách con người nhận thức về độ tương tự. Mỗi lớp biểu diễn một chủ đề ngữ nghĩa khác nhau, các ảnh trong cùng một lớp được xem là liên quan.

Bảng 1.*Các tập ảnh.*

Tập ảnh	Số lượng ảnh	Số lớp ảnh
Logo-2K+	22725	303
VGGFACE2-S	60000	500

Tập dữ liệu thứ nhất Logo-2K+[12], gồm 22725 hình ảnh logo của 303 thương hiệu khác nhau và được đặt trong 303 nhóm, với mỗi nhóm gồm 75 ảnh của mỗi thương hiệu. Hình 1 chỉ ra một số mẫu ảnh trong tập dữ liệu này.



Hình 1. *Một số mẫu trong tập dữ liệu ảnh Logo-2K+.*

Tập dữ liệu thứ 2 VGGFACE2-S là tập con của tập dữ liệu hình ảnh để nhận dạng khuôn mặt VGGFACE2, bao gồm 60000 ảnh chia thành 500 nhóm với mỗi nhóm 120 ảnh của mỗi người. Hình 2 là một số hình ảnh trong tập dữ liệu này (tập dữ liệu này được lấy ngẫu nhiên từ tập dữ liệu VGGFACE2với 169396 ảnh trong 500 lớp [13]). Hình ảnh trong các tập dữ liệu có kích thước và mầu sắc khác nhau.



Hình 2. *Một số mẫu trong tập dữ liệu ảnh VGGFACE2-S.*

4.1. Trích chọn đặc trưng

Ảnh mầu hoặc ảnh đa cấp xám được biểu diễn bằng các đặc trưng mức thấp thuộc nhiều bộ (bộ mô tả về mầu, bộ mô tả về kết cấu hoặc mô tả về hình dạng,…).

Trong các thực nghiệm, chúng tôi sử dụng 5 đặc trưng toàn cục để mô tả một ảnh: Color Moments, LBP, Gabor Wavelets Texture, Edge và GIST (xem [2, 3, 7]).

Khoảng cách Euclid sẽ được sử dụng để tính khoảng cách giữa các đặc trưng. Các đặc trưng mức thấp được chuẩn hóa bằng phương pháp 3σ-opt để mỗi thành phần của mỗi vector đặc trưng mức thấp của cơ sở dữ liệu ảnh nằm trong khoảng [-1,1] (xem [14]).

4.2. Các kết quả và luận giải

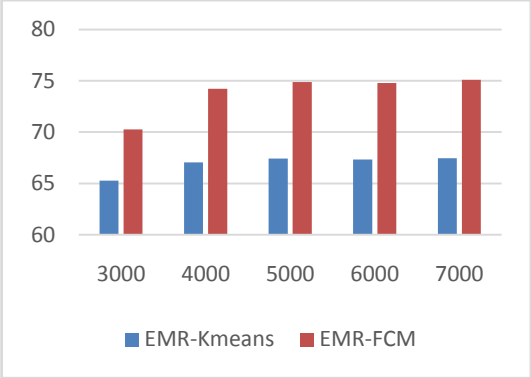
Trong thực nghiệm này, chúng tôi chọn $\mu_\varepsilon = 10^{-6}$, nb = 500 đối với thuật toán FCM đề xuất. Để đánh giá khách quan hiệu quả của thuật toán EMR-Kmeans và EMR-FCM đề xuất, chúng tôi sử dụng một chỉ số tương tự độ đo Average Precision (chúng tôi vẫn gọi là AP) được đề xuất bởi NISTTREC video (TRECVID) [14], AP được định nghĩa trung bình của giá trị độ chính xác thu được sau mỗi ảnh liên quan được tra cứu.

Tập ảnh truy vấn Q được chọn ngẫu nhiên với số lượng 10% theo từng chủ đề của tập ảnh thử nghiệm Logo-2k+ và VGGFACE2-S.

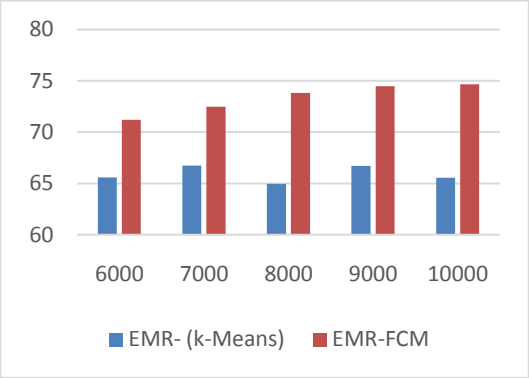
Với mỗi ảnh truy vấn $q \in Q$, sử dụng các độ đo tương tự cho bởi EMR-Kmean và EMR- FCM đề xuất, chúng ta chọn N = 100 ảnh có độ tương tự cao nhất. Giá trị độ chính xác là trung bình tỷ lệ giữa số ảnh liên quan trong N ảnh được trả lại bởi các giá trị tương tự với từng ảnh q. Gọi tập các phần tử liên quan đến truy vấn $q \in Q$ là $\{d_1, d_2, ..., d_{mj}\}$, giá trị AP trên toàn bộ các truy vấn được tính toán như sau:

$$AP(Q) = \left(\frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{m_j}{N} \right) * 100 \qquad (13)$$

Từ các thử nghiệm trên các tập dữ liệu khác nhau (Logo-2k+và VGGFACE2-S), tùy thuộc vào số hình ảnh, lớp hình ảnh và hình ảnh trong mỗi lớp việc lựa chọn các tham số C (điểm neo – anchor point) khác nhau sẽ cho kết quả khác nhau. Và thuật toán đề xuấtEMR sử dụng FCM để chọn các điểm neo (chính là tâm của các cụm thu nhận được) cho kết quả tốt hơn so với thuật toán EMR gốc sử dụng K-means.



Kết quả thử nghiệm trên cơ sở dữ liệu Logo-2K+



Kết quả thử nghiệm trên cơ sở dữ liệu VGGFACE2-S

Hình 3. *Kết quả thực nghiệm trên các tập dữ liệu.*

Hiệu quả truy vấn trung bình của EMR trên 2 tập dataset (Logo-2k+ và VGGFACE2-S) là 66.91%; 65.90%. Hiệu quả truy vấn trung bình của EMR cải tiến trên 2 tập dataset (Logo-2k+ và VGGFACE2-S) là 73.85%; 73.33%.

5. KẾT LUẬN

Trong bài báo này, thuật toán FCM đã được cải tiến đề xuất để phân cụm một cơ sở dữ liệu lớn cho các vectơ nhiều chiều thành nhiều cụm. Sau đó, thay vì sử dụng K-means như thông thường, FCM được áp dụng cho bước chọn điểm neo của EMR. Do đó, cho ra một EMR tốt hơn, cụ thể là EMR-FCM cải tiến và độ chính xác của việc truy xuất hình ảnh bằng EMR đề xuất cao hơn so với thuật toán gốc. Đối với cơ sở dữ liệu hình ảnh lớn, các hệ thống CBIR trích xuất các đặc trưng thấp của ảnh trong pha ngoại tuyến và xếp hạng các vectơ đặc trưng hình ảnh theo mức độ tương đương với vectơ đặc trưng hình ảnh truy vấn đã được xây dựng bằng cách sử dụng EMR- FCM trong pha trực tuyến. Với cơ sở dữ liệu hình ảnh lớn, thời gian tính toán không bị ảnh hưởng vì hầu hết các mô-đun tính toán như trích xuất đặc trưng cấp thấp, các mô-đun liên quan FCM cải tiến áp dụng cho toàn bộ cơ sở dữ liệu đã được giải quyết trong giai đoạn ngoại tuyến. Kết quả thực nghiệm là thách thức thúc đẩy chúng tôi thử nghiệm trên hình ảnh y tế lớn với hàng ngàn hình ảnh được thêm vào cơ sở dữ liệu được lưu trữ bằng nhiều cách khác nhau.

TÀI LIỆU THAM KHẢO

[1]. Bin Xu, Jiajun Bu, Chun Chen, Can Wang, Deng Cai and Xiaofei He. EMR: *"A Scalable Graph-based Ranking Model for Content-based Image Retrieval"*, IEEE transactions on knowledge and data engineering. **27**, no.1, 102-114 (2015).

[2]. Wang, M., Fu, W., Hao, S., Tao, D., & Wu, X. *"Scalable Semi-Supervised Learning by Efficient Anchor Graph Regularization"*. IEEE Transactions on Knowledge and Data Engineering. **28** (7), 1864–1877 (2016).

[3]. Liang, S., Markov, I., Ren, Z., & de Rijke. M. *"Manifold Learning for Rank Aggregation"*. Proceedings of the 2018 World Wide Web. Conference on World Wide Web - WWW '18. 2018.

[4]. Wu, Y., Wang, X., & Zhang, T. *"Crime Scene Shoeprint Retrieval Using Hybrid Features and Neighboring Images"*. Information, **10** (2), 45 (2019).
[5]. Shaoyan Sun, Ying Li, Wengang Zhou, Qi Tian c, Houqiang Li. *"Local residual similarity for image re-ranking",* Information Sciences.**417**, 143–153(2017).
[6]. Anil K. Jain. *"Data clustering: 50 years beyond K-means",* Pattern Recognition Letters.**31,** 651–666 (2010).
[7]. J.C. Bezdek, *"Pattern Recognition with Fuzzy Objective Function Algorithm",*Plenum Press, 1981.
[8]. Yang, Miin-Shen, Pei-Yuan Hwang, and De-Hua Chen. *"Fuzzy clusterinsg algorithms for mixed feature variables"*. Fuzzy Sets and Systems. **141**,no. 2. 301-317 (2004).
[9]. Hanuman Verma, Akshansh Gupta, Dhirendra Kumar. *"A modified intuitionistic fuzzy c-means algorithm incorporating hesitation degree",* Pattern Recognition Letters.**122**, 45–52.(2019).
[10]. Timothy C. Havens, James C. Bezdek, Christopher Leckie, Lawrence O. Hall, and Marimuthu Palaniswami. *"Fuzzy c-Means Algorithms for Very Large Data",* IEEE Transactions on Fuzzy systems. **20**, no.6, December 2012.
[11]. Kuo-Lung Wu. *"Analysis of parameter selections for fuzzy c-means",* Pattern Recognition.**45,**407–415 (2012).
[12]. Dataset. [Online]. Available: https://drive.google.com/drive/folders/1PTA24UTZcsnzXPN1gmV0_lRg3lMHqwp6
[13]. Dataset. [Online] http://zeus.robots.ox.ac.uk/vgg_face2/
[14]. Trung Hoang Xuan, Tuyet Dao Van, Huy Ngo Hoang, Sergey Ablameyko, Cuong Nguyen Quoc, Quy Hoang Van. *"A Novel Non-Gaussian Feature Normalization Method and itsApplication in Content Based Image Retrieval"*. Nonlinear Phenomena in Complex Systems.**22**, no. 1, pp. 1 – 17 (2019).

ABSTRACT

MANIFOLD RANKING ON MULTIPLE LOW-LEVEL FEATURE SET NORMALIZED WITH OPTIMIZED PARAMETERS IN CBIR

In CBIR, the image is represented by multi low-level features that describe the color, texture and shape of the image. The combination of different image features in global similarity measurements such as the EMR requires normalized data sets. In this paper, a new normalization method for vector number data such as the low level features of color images is proposed. Experimentation has shown the effectiveness of the proposed algorithm for the manifold ranking EMR, and the CBIR quality is really improved.

Keywords: Big data; EMR; K-means; FCM; CBIR.

Nhận bài ngày 27 tháng 5 năm 2020
Hoàn thiện ngày 10 tháng 7 năm 2020
Chấp nhận đăng ngày 15 tháng 10 năm 2020

Địa chỉ: [1]Đại học Hồng Đức, Thanh Hóa;
[2]Đại học Điện lực;
[3]Đại học Kinh doanh và Công nghệ Hà Nội;
[4]Trường Đại học Hải Phòng.
[*]Email: minhquy1208@gmail.com.