DOI: https://doi.org/10.31341/jios.42.2.4 UDC 004.932
Open Access Original Scientific Paper

An Efficient and Effective Image Retrieval System on the basis of Feature, Matching Measure and sub-space Selection

Mawloud Mosbah *mos_nasa@hotmail.fr*
Informatics Department
Faculty of Sciences
University 20 Août 1955 of Skikda, Algeria

Abstract

Since its appearance as a research field, Content-based Image Retrieval (CBIR) system has increasingly received an important attention. Review of literature reveals that the efforts put, up to now, in the field address either effectiveness or efficiency. In this paper, we address both accuracy and efficiencythrough introducing an efficient and an effective image retrieval approach based on feature, matching measure and sub-spaceselection. The selection relies on relevance feedback information injected by the user. The approach is tested on Corel-1Kimages database. The obtained results are very promising.

Keywords: Content Based Image Retrieval, Feature Selection, Matching Measure Selection, Sub-space Selection, efficiency, effectiveness.

1. Introduction

Unlike database system, handling structured data, the information retrieval one aims to retrieve information from a colossal unstructured and semi-structured collection. No matter what information being retrieved: text, image or video, an information retrieval system keeps the same architecture composed of three components: indexing stage, matching process and interrogation protocol. As the user is not satisfied from the returned results, it is required to improve the information retrieval system. This improvement implies then to improve the pre-cited components. Indeed, many efforts have been put, for the case of the image, in the last decades, in order to enhance the effectiveness of mage retrieval system [1], [2] through enhancing its components [3], [4], [5], [6]. Efficiency [7] is also taken into account for getting a real time system, although the big volume of databases being retrieved.

The key question to be asked is what feature and what matching measure being considered for building a CBIR system [8] of high quality? The answer for such question is selection paradigm either within indexing stage through feature selection [9] or at the matching process via matching measure selection [6].The both kinds of selection are based on relevance feedback information injected by the user.

For the efficiency aspect, a sub-space selection guided by feature and matching measure selection seems to be a good idea. Indeed, the organization of the collection as clusters or trees makes the retrieval operation efficient. Moreover, sub-space selection may contribute into the improvement of the system performance.

In this paper, we introduce an efficient and an effective image retrieval approach based on Feature, matching measure and sub-space selection.

The rest of the paper is arranged as follows: Section 1 deals with feature selection. In Section 2, we talk about matching measure selection. Section 3 addresses relevance feedback mechanism. In Section 4, we introduce our approach that of sub-space selection on the basis of feature and matching measure selection. We conclude the paper with a conclusion.

2. Feature Selection

Feature selection refers to as choosing the features ‘combination among a given large set that well describes a particular data collection [10]. The purpose of feature selection mechanism, applied into a large spectrum of fields such as pattern recognition and data mining, consists of designating the discrimination power of features and tackling the dimensionality curse [12]. For CBIR field, this mechanism is applied for encountering the semantic gap [10], [11], [12], [13], [14].

In [9], authors have optimized feature selection parameters to reach a maximum precision of CBIR systems. In [11], authors have proposed two feature selection criteria based on inner-cluster and inter-cluster relations. In [12], Jiang et al have proposed a feature selection criterion based on computing similarity between the relevant and irrelevant image sets and an effective online feature selection algorithm. In [13], Lu et al have introduced a novel method baptized Principal Feature Analysis (PFA). This method proceeds to choose the principal features in face tracking and CBIR problems. In [14], Benloucif and Boucheham have conducted a comparative study of Greedy Heuristic, Tabu Search and Genetic Algorithms and their impact on the performance of CBIR.

3. Matching Measure Selection

Review of literature reveals some efforts put into matching process such as [3], [15], [16], [17], [18]. In [15], authors have provided a systematic comparison of various similarity measures in the medical CBIR application context. In [16], Perlibakas has compared 14 distance measures and their modifications between feature vectors with respect to recognition performance of the principal component analysis (PCA)-based face recognition method. In [17], a comprehensive performance study has been conducted for sixteen dissimilarity measures, on seven typical feature spaces, using two search methods. In [18], different similarity measurements, commonly used in image retrieval, have been described and evaluated using shape features and standard shape datasets. In [19], Cha has enumerated and categorized a large variety of distance/similarity measures for comparing nominal type histograms. In [7], we have conducted a comparison between many matching measures (distances, quasi-distances, similarities and divergences), in the context of CBIR, in terms of effectiveness and efficiency. In [6], we have introduced a new a paradigm that of matching measure selection. The study has considered as many as 18 matching measures, including similarities, distances, quasi-distances and divergences. The selection process is based on the SFS algorithm with one round and relevance feedback for determining the best matching measure for a specific query. The obtained results show that the proposed approach that of matching measure selection yields promising results in terms of precision, recall [20] and utility value [21].

4. The Proposed Approach: Sub-Space Selection

For taking into account the efficiency aspect, we have to organize the indexing space in a way allowing a direct access for relevant information. This organization may rely on some structures such as clusters, B-trees [22], R-trees [23] and X-trees [24]. In this work, we consider the clustering alternative using k-means algorithm [25]. The proposed approach is depicted in Figure1.

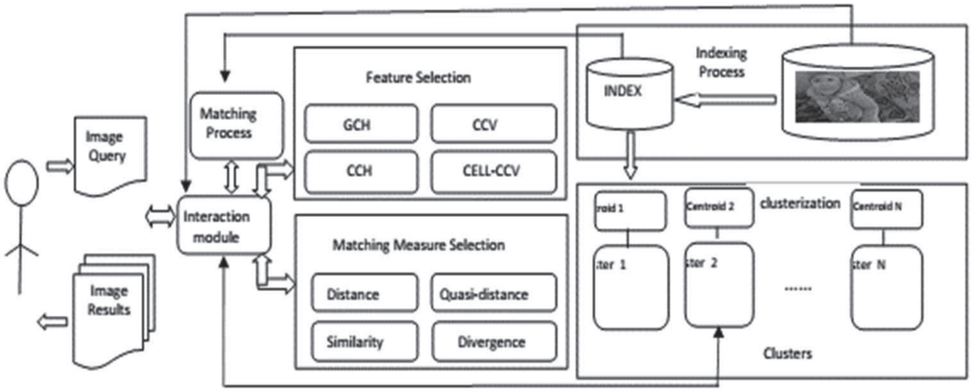

Figure 1. The general architecture of the proposed approach.

Given that indexing space is organized as clusters and each cluster is represented by each centroid, the execution scenario of the introduced approach can be described as follows: after that the user submits the query, the system answers by a set of images as initial results considering one feature and one matching measure. After that, the user has to inject his/her relevance feedback information through designating some relevant images from those answered by the system. The fist images not labelled by the user are assumed as non relevant. The system will then run a selection algorithmsuch as the SFS with one round applied in [6] which designates the best feature and the best matching measure.The submitted query encoded by the selected feature will be compared to clusters centroids employing the selected matching measure. The images of the cluster that the centroid is the closest to the query will be visualized to the user as new results.

The indexing space is arranged then as depicted in Figure 2.

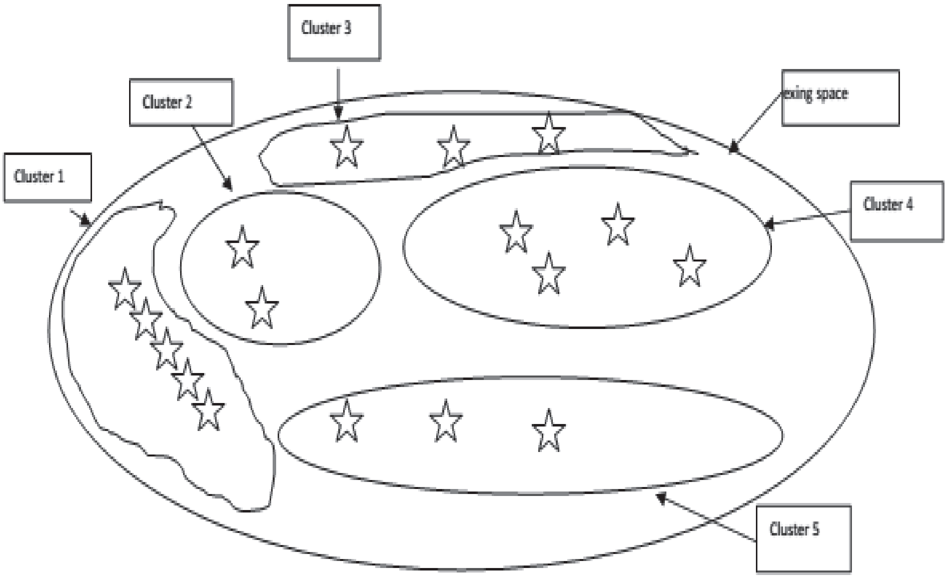

Figure 2. Indexing space decomposed into different clusters.

Even effectiveness is taken into consideration by the proposed approach through feature selection and matching measure selection.

The pseudo code for selecting the best configuration (feature, matching measure) is given as follows:

```
Step1: as initialization, the algorithm starts with the following weighting (0, 0,.., 0) for the features and matching measures. (Neither selected feature nor selected matching measure).
Step2: each weight will be set to 1 separately to generate many configurations.
Step3: evaluating each configuration (feature, matching measure) based on fitness.
Step4: selecting the best configuration (feature, matching measure).
Step5: comparing the actual selected configuration with the selected configuration (feature, matching measure) and go to step8.
Step6: set the other weights to 0 except the weight of the selected (feature, matching measure) is still 1.
Step7: go to step 2.
Step8: END.
```

Unlike the work done in [6], this work implements SFS completely through combining matching measures. For matching measures evaluation, all matching measures (similarity, distance, quasi distance and divergence) should be converted to similarity then added together for getting one value. The conversion from the other matching measures to similarity measure is materialized as follows:

- For *distance D*: *1-D*.
- For *quasi-distance QD* (and for *Divergence DV*): *1-normalisation (QD)*. For normalisation: $\frac{QD}{\max(QD)}$

The pseudo code for selecting the best*sub-space* is given as follows:

```
Step 1: receiving Relevance Feedback from the user.

Step 2: selecting configuration (Feature, matching measure) according to the user
        relevance feedback.

Step 3: adopting selected configuration for computing similarity between the
        submitted query and the centroid of each cluster.

Step 4: visualizing the selected cluster as new outputs for the user.

END.
```

5. Materials and Experiments

In this Sub-Section, we test the effectiveness and the efficiency of the considered sub-space selection approach on the basis of feature and matching measure selection. The experiments are conducted on COREL-1K benchmark [26]. As signature, we have used Global Color Histogram (GCH) [27], Color Coherence Vector (CCV) [28], Cell-ColorHistogam (CCH) [29] and CELL-CCV[30]. For matching measure, we have utilized (Euclidean, Manhattan, Intersection, Sorensen, Kulczunsky, Soergel, Chebyshev, Squared, Mahalanobis and Canberra) distances, (Ruzicka, Roberts, Motyka and Cosine) similarities, (X^2, Neyman-X^2 and Separation) quasi-distances and Jeffreydivergence [7].

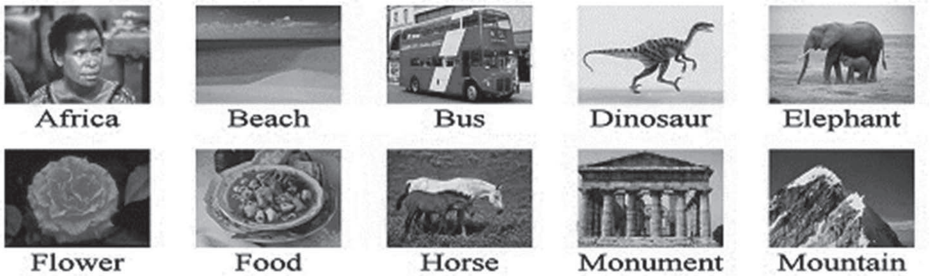

Figure 3. The Semantic Class of COREL-1K Images Collection.

Table 1 shows the average weighting precision of different considered indexing methods, feature selection case, matching measure selection case and the sub-space selection case.

AWP on	Average Weighted Precision (%)							
	GCH (using Man) [1]	CCV (using Man) [1]	CCH (using Man) [1]	CELL-CCV (using Man) [1]	Feature Selection (using Man)	MMS(using CCV)	SSS based on FS and MMS	SSS based on the query, Man, CCV
C1	45	45.8	42.70	47	47	50	50	45.8
C2	30.80	28.10	34.20	31.20	31.20	20	31.20	28.10
C3	33.50	34.50	33.40	34	34	20	34	34.50
C4	42.24	44.40	39.30	45	45	40	45	44.40
C5	95.75	95.90	97.30	97.90	97.90	100	100	95.90
C6	50.10	52.80	53.90	56.40	56.40	70	70	52.80
C7	70.10	77.70	73.50	83.60	83.60	80	83.60	77.70
C8	84.70	82.70	83.90	83.70	84.70	20	84.70	82.70
C9	47.90	41.20	45.30	44.60	47.90	40	47.90	41.20
C10	49.60	48.60	49.80	51.40	51.40	60	60	48.60
Aver	54.96	55.17	55.54	57.48	57.91	50	60.64	55.17
CT (ms)	T1=46.8	T2=63.2	T3=72.1	T4=76	T5	T6	T7	T8

Table 1. Effectiveness and Efficiency over Considered Features,Feature Selection, Matching Measure Selection and Sub-Space Selection.

Where AWP is the Average Weighted Precision, Man is Manhattan distance, MMS is Matching Measure Selection, FS is Feature Selection, SSS is Sub-Space Selection, (C1, C2, C3, C4, C5, C6, C7, C8, C9, C10) are respectively the classes (People, beach, building, bus, dinosaur, elephant, rose, horse, mountain, food), Aver is the Average, CT is consumed time.

As we utilized only 10 images as relevance feedback, T5, T6, T7and T8 are given as follows:

$$T5 = \begin{cases} 49.381\ or \\ 65.781\ or \\ 74.681\ or \\ 78.581 \end{cases}$$

$$T6 = \begin{cases} 21.372\ or \\ 36.972\ or \\ 52.572\ or \\ 68.172 \end{cases}$$

$$T7 = \begin{cases} 24.577\ or \\ 24.733 or \\ 24.889\ or \\ 25.045 \end{cases}$$

$$T8 = \begin{cases} 23.666\ or \\ 23.816 or \\ 23.967\ or \\ 24.117 \end{cases}$$

To note that, the consumed time includes indexing query image time, time of computing similarities with the images collection and ranking time.

As shown in results above, sub-space selection performance outperforms the performance in the case using the different considered signatures and matching measure and even outperforms the quality in the case of feature selection and matching measure selection. For the efficiency aspect, there is a great superiority in the case when adopting sub-space selection comparing to other cases.

6. Conclusion

In this paper, we have introduced a novel approach which takes into account both effectiveness and efficiency when retrieving images. The approach proceeds to organise the images collection as clusters and selecting the best configuration (feature, matching measure) on the basis of relevance feedback information. Selected feature and matching measure are utilized for designating the cluster to be visualized to the user. The experiments conducted on COREL-1K indicate that the approach is so promising.

References

[1] Y. Rui, T. S. Huang and S. F. Chang, “Image retrieval: Current techniques, promising directions, and open issues”. *Journal of visual communication and image representation*, vol. 10, no. 1, pp. 39-62, 1999.

[2] R. Datta, D. Joshi, J. Li and J. Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age”. *ACM Computing Surveys (Csur)*, vol. 40, no. 2, pp. 5, 2008.

[3] M. Mosbah, “Mesures de distance dans le cadre de la recherche d’images par le contenu (CBIR) ”, Doctorate thesis, University 20 Août 1955 of Skikda, 2017.

[4] M. Salmi and B. Boucheham, “Content based image retrieval based on cell color coherence vector (Cell-CCV)”. In Proc. The 4th International Symposium: Concepts and Tools for Knowledge Management (ISKO-Maghreb), ’11, 2014, pp. 1-5.

[5] M. Mosbah and B. Boucheham, “Relevance feedback within CBIR systems”. *International Journal of Computer, Information Science and Engineering*, vol. 8, no. 4, pp. 19-23, 2014.

[6] M. Mosbah and B. Boucheham, “Distance selection based on relevance feedback in the context of CBIR using the SFS meta-heuristic with one round”. *Egyptian Informatics Journal*, vol. 18, no. 1, pp. 1-9, 2017.

[7] M. Mosbah and B. Boucheham, (2017, April). “Matching Measures in the Context of CBIR: A Comparative Study in Terms of Effectiveness and Efficiency”. In Proc. World Conference on Information Systems and Technologies, ’04, 2017 pp. 245-258.

[8] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, A. and R. Jain, “Content-based image retrieval at the end of the early years”. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 12, pp. 1349-1380, 2000.

[9] J. G. Dy,C. E. Brodley, A. Kak, L. S. Broderick and A. M. Aisen, “Unsupervised feature selection applied to content-based retrieval of lung images”. *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 3, pp. 373-378, 2003.

[10] E. Guldogan and M. Gabbouj “Feature selection for content-based image retrieval”. *Signal, Image and Video Processing*, vol. 2, no. 3, pp. 241-250, 2008.

[11] E. Rashedi, H. Nezamabadi-Pour and S. Saryazdi “A simultaneous feature adaptation and feature selection method for content-based image retrieval systems”. *Knowledge-Based Systems*, vol. 39, pp. 85-94, 2013.

[12] W. Jiang, G. Er, Q. Dai, and J. Gu “Similarity-based online feature selection in content-based image retrieval”. *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 702-712, 2006.

[13] Y. Lu, I. Cohen, X. S. Zhou and Q. Tian, “Feature selection using principal feature analysis”. In Proc. The 15th ACM international conference on Multimedia, ’09, 2007, pp. 301-304.

[14] S. Benloucif and B. Boucheham, “Impact of feature selection on the performance of content-based image retrieval (CBIR)”. In Proc. The 4th International Symposium: Concepts and Tools for Knowledge Management (ISKO-Maghreb), ’11, 2014, pp. 1-7.

[15] J. Collins and K. Okada “A Comparative Study of Similarity Measures for Content-Based Medical Image Retrieval”. In CLEF (Online Working Notes/Labs/Workshop), Sept. 2012.

[16] V. Perlibakas, “Distance measures for PCA-based face recognition”. *Pattern recognition letters*, vol. 25, no. 6, p. 711-724, 2004.

[17] R. Hu, S. Ruger, D. Song, H. Liu and Z. Huang, “Dissimilarity measures for content-based image retrieval”. In Proc. International Conference on Multimedia and Expo, ’06 2008, pp. 1365-1368.

[18] D. Zhang and G. Lu, “Evaluation of similarity measurement for image retrieval”. In Proc. International Conference on Neural Networks and Signal Processing, ’12, 2003, pp. 928-931.

[19] S. H. Cha, “Comprehensive survey on distance/similarity measures between probability density functions”. *City*, vol. 1, no. 2, 2007.

[20] G. P. Babu, B. M. Mehtre and M. S. Kankanhalli, “Color indexing for efficient image retrieval”. *Multimedia Tools and applications*, vol. 1, no. 4, p. 327-348, 1995.

[21] P. Fishburn, “Non-linear preference and utility theory”. Johns Hopkins University Press, 1998.

[22] V. Srinivasan and M. J. Carey, *Performance of B-tree concurrency control algorithms*, New-York: ACM, 1991, vol. 20, no. 2, pp. 416-425.

[23] A. Guttman, *R-trees: A dynamic index structure for spatial searching*, New-York: ACM, 1984, vol. 14, no. 2, pp. 47-57.

[24] S. Berchtold, D. A. Keim and H. P. Kriegel, “An index structure for high-dimensional data”. In Proc. The 22th International Conference on Very Large Data-Bases, ’09, 2001, pp. 28-39.

[25] M. Mosbah and B. Boucheham, “Re-ranking in the Context of CBIR: A Comparative Study”. In Proc. World Conference on Information Systems and Technologies, ’04, 2017, pp. 297-307.

[26] Wang Group: Modelling Objects, Concepts, Aesthetics, and Emotions in Big Visual Data", Wang.ist.psu.edu, 2018. [Online]. Available: http://Wang.ist.psu.edu/docs/home.shtml [Accessed March 9, 2014].

[27] M. J. Swain and D. H. Ballard, “Color indexing”. *International journal of computer vision*, vol. 7, no. 1, p. 11-32, 1991.

[28] G. Pass, R. Zabih and J. Miller, “Comparing images using color coherence vectors”. In Proc. The 4th ACM international conference on Multimedia, ’02, 1997, pp. 65-73.

[29] R. O. Stehling, M. A. Nascimento and A. X. Falcão, “Cell histograms versus color histograms for image representation and retrieval”. *Knowledge and Information Systems*, vol. 5, no. 3, p. 315-336, 2003.