Scientific Research Publishing

Journal of Computer and Communications, 2022, 10, 177-185
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Survey on Clustering Techniques for Image Categorization Dataset

Mohd Afizi Mohd Shukran[1*], Mohd Sidek Fadhil Mohd Yunus[1], Muhammad Naim Abdullah[2], Mohd Rizal Mohd Isa[1], Mohammad Adib Khairuddin[1], Kamaruzaman Maskat[1], Suhaila Ismail[1], Abdul Samad Shibghatullah[3]

[1]Universiti Pertahanan Nasional Malaysia, Kuala Lumpur, Malaysia
[2]University Malaysia of Computer Science & Engineering (UNIMY), Cyberjaya, Malaysia
[3]UCSI University, Kuala Lumpur, Malaysia
Email: *afizi@upnm.edu.my, sidek@upnm.edu.my, naim.abdullah@unimy.edu.my,rizal@upnm.edu.my, adib@upnm.edu.my, kamaruzaman@upnm.edu.my, suhaila@upnm.edu.my, abdulsamad@ucsiuniversity.edu.my

How to cite this paper: Shukran, M.A.M., Yunus, M.S.F.M., Abdullah, M.N., Isa, M.R.M., Khairuddin, M.A., Maskat, K., Ismail, S. and Shibghatullah, A.S. (2022) Survey on Clustering Techniques for Image Categorization Dataset. *Journal of Computer and Communications*, **10**, 177-185.
https://doi.org/10.4236/jcc.2022.106014

Received: June 5, 2022
Accepted: June 27, 2022
Published: June 30, 2022

Abstract

Content Based Image Retrieval, CBIR, performed an automated classification task for a queried image. It could relieve a user from the laborious and time-consuming metadata assigning for an image while working on massive image collection. For an image, user's definition or description is subjective where it could belong to different categories as defined by different users. Human based categorization and computer-based categorization might produce different results due to different categorization criteria that rely on dataset structure and the clustering techniques. This paper is aimed to exhibit an idea for planning the dataset structure and choosing the clustering algorithm for CBIR implementation. There are 5 sections arranged in this paper; CBIR and QBE concepts are introduced in Section 1, related image categorization research is listed in Section 2, the 5 type of image clustering are described in Section 3, comparative analysis in Section 4, and Section 5 conclude this study. Outcome of this paper will be benefiting CBIR developer for various applications.

Keywords

Categorization, CBIR, Classifications, Clustering, Dataset

1. Introduction

In CBIR and human interaction, human level image interpretation is applicable where a human user that performed the image query expected CBIR to produce a result based on their preferences. Human based image interpretation is depending on users' psychological and knowledge level [1] while computer-based classification is relying on clustering techniques. Computer based image classification can be divided into 3 modules; extraction module, query module and retrieval module [2]. An image is interpreted by computer during query module where extracted feature is linked with suitable metadata such as keywords, captioning, or hash-tag (a popular manual metadata technique among netizen lately).

Clustering techniques act to determine the suitable and related metadata for a queried image by referring the similarity matrix of vector data group in a dataset. The referring dataset also called as training dataset, contained number of vector data (extracted feature of an image) that was divided into few data groups with each group carrying different data property (such as image's keyword) using a clustering technique. Clustering also is a self-learning way for CBIR to find the queried image similarity from the image dataset. The method of querying an image to find a set of similar images is known as Query by Example, QBE [3] where CBIR will return results in the form as image similarity or defined keywords from the user's loaded image.

However, the QBE style query is fully relying on data contained in the training dataset that results accuracy probability might vary. The outcome of QBE style query accuracy might become deficient when the sum of data in training dataset is insufficient or the dataset improperly cluster.

2. Related Research

While most research focusing on clustering large number of data such in [4] research and multi-cluster data such as [5] research, there are just a few researchers focusing on complex structure data as in [6] research. Human level image categorization is complex when a CBIR deal with a lot of users' interpretation. Well, CBIR is created to serve human, CBIR need to fulfilling human interpretation. [7] is crowd-source image dataset that categorized images based on English dictionary tree structure which contained around 20 thousand categories where hundreds of images reside in each category (**Figure 1**).

Practically, other than public hosted CBIR such as [8] image, most of private CBIR did not fully utilized or require such 20 thousand category datasets. For a specific purposed CBIR system, it is more reliable to containing only the purposed related dataset. For instance, a car finder CBIR system only utilized car related dataset should be enough. Other than that enormous dictionary-based category dataset, there are few other research purpose image datasets provided from various academic and research institutional. Those datasets provided only specific purposed content with a smaller number of categories. The only thing to consider is the type of dataset that adequate for a CBIR (**Table 1**).

There more than that listed dataset available to use as a sample for a research such as surveillance camera, animals, nature process and much more. But, above of all, the purpose of a CBIR system development is a critical decision that affected the content of dataset to adopt.

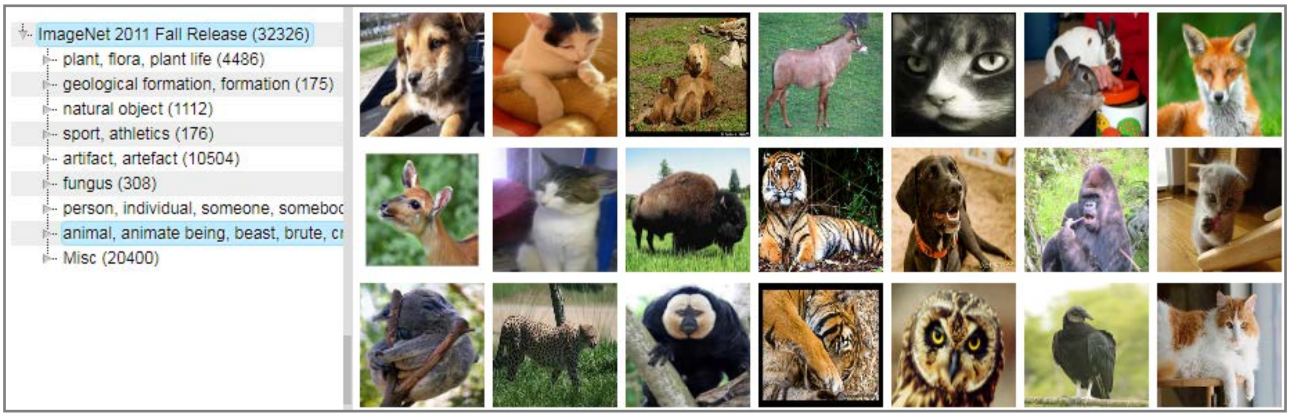

ImageNet 2011 Fall Release (32326)
plant, flora, plant life (4486)
geological formation, formation (175)
natural object (1112)
sport, athletics (176)
artifact, artefact (10504)
fungus (308)
person, individual, someone, somebod
animal, animate being, beast, brute, cr
Misc (20400)

Figure 1. A part of ImageNet hierarchical dataset screenshot.

Table 1. Example list of specific content research dataset.

Dataset category	Dataset name	Content Description	Source/Creator/Owner
Facial Recognition	Grammatical Facial Expressions Dataset	• 27,965 images extracted from Microsoft Kinect feature of common facial expression • Text form dataset that ready for clustering	[9]
	IMDB-WIKI	• 523,051 faces collected from IMDB and Wikipedia image collection • Mainly suggested for gender and age classification assignment • Image dataset need to be extracted	[10]
Body posed	Activitynet	• 10,024 labeled articles in the form of images and videos dataset for action detection • Articles need to be extracted and clustered with the label as a reference	[11]
Objects	CIFAR-100 Dataset	• 60,000 images of objects that segmented into 100 categories • Image dataset for feature extraction and clustering	[12]
	Fashion MNIST	• 60,000 images of fashions dataset for feature extraction and clustering	[13]
Alphanumeric recognition	Gisette Dataset	• 13,500 images of confusing handwriting/bad handwriting that consist of 4 to 9 characters • Image dataset for feature extraction and clustering	[14]
	MNIST Database	• 60,000 images of handwritten digits • Image dataset for feature extraction and clustering with label	[15]
Area, Places, Zones	LabelMe	• 187,240 images of scene at various location for example in the classroom, at coffee bar or seaside • Image dataset for feature extraction and clustering with label	[16]
	SpaceNet	• More than 17,500 images from commercial satellite that also containing building footprint • Image dataset for feature extraction and clustering with label	[17]

3. The Clustering Techniques

CBIR employ the clustering techniques perform the data categorization that reside in a dataset. Clustering techniques is consisting of a collection of clustering algorithms which commonly divided into 5 categories [18] based on their clustering behavioral. Clustering algorithm is a self-learn process for CBIR to link a suitable metadata captions and keywords for a queried image. On a preloaded dataset (or adopted dataset), clustering algorithm analyzing the dataset in matrices (either mostly 2-dimension matrix or some 3-dimension matrix) to find the clustering centers among of the vector data (**Figure 2**).

The cluster area is built around the cluster center and assigning each data that resides in a cluster area as a cluster membership. Some clustering algorithm is assigning the cluster membership employing the hard-clustering rule where each data strictly become a member of one cluster only. While the other clustering rule is a data could become one or more cluster membership (soft clustering). Hard clustering is less complex than soft clustering and thus, in term of performance it performs faster with minimal resources consumption. However, in several CBIR application, an image is required to allow residing to more than just one category and that make soft clustering rule cannot be abandon.

During the query process, the clustering algorithms calculate the similarity matrices between extracted feature of queried image and the data in a dataset. If the newly queried data need to be inserting into dataset matrix, the clustering algorithm will be revising the cluster center and as well as cluster area for any changes (the term iteration is used for partition-based clustering algorithm). There will be possibilities where a cluster is expanding, downsizing or a new cluster creation. From **Table 2** previously, it shows that each clustering algorithm perform the data clustering with their own ways and produced different result each. The range of clustering algorithms variety is needed for many purposes, reasons, and requirements of a CBIR development.

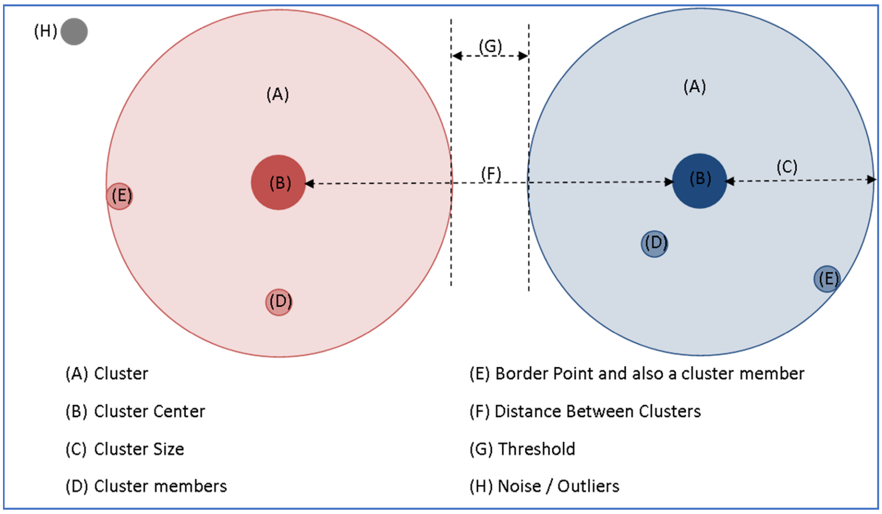


Figure 2. The clusters' notions.

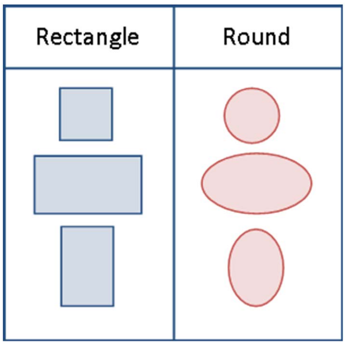
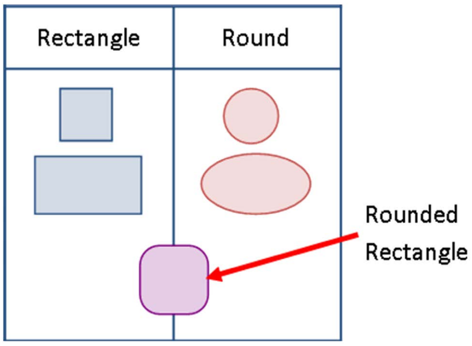
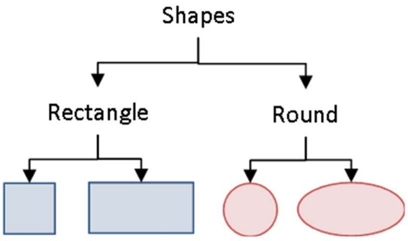
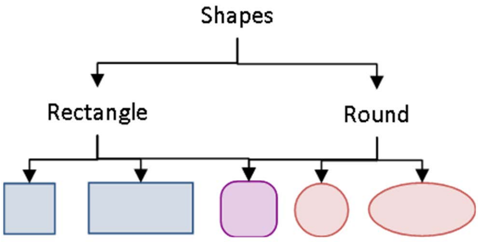
Table 2. List of clustering techniques.

Techniques	Description	Algorithms' brief description
Density and Model based	• Cluster center is built on the density area, a hotspot where data most plotted area • Cluster is built around the cluster center and cluster size could growing where there is threshold • If there is a new density area, a new cluster will form • For all density and model-based clustering techniques algorithm, it shows deficiency when density area is wide, and form a non-rounded hotspot such as L shape	**DBSCAN [19]** • Hard clustering, need to set the minimum number of densities • Filter out noise, but, if the minimum density value is too high, a useful data would become noise • If the density value is set too low, a lot of clusters is formed and would slower the query process **Expectation Maximization, EM [20]** • Soft clustering based on Gaussian peak density that could overlap on another cluster area
Grid based	• Logical grid is build based on maximum grid size (Cluster Size) • Less time consumption on cluster exploration • However, a data might be placed on a wrong cluster since grids is fixed, hard clustering method	
Hierarchical based	• There are 2 ways cluster is form, agglomerative and divisive • Agglomerative is merging few data to form a cluster and merging few clusters to form a greater cluster • Divisive is splitting a large cluster into few sub cluster • Agglomerative is complex to form but fast to explore while divisive is vice versa	**HDBSCAN [21]** • While the other hierarchical based algorithm was built based on distance linkage, this algorithm is build based on density • Distance linkage method, once merging or splitting performed, it is irreversible • But, compared to the other hierarchical techniques, HDBSCAN face performance issues as the other density-based techniques
Partition based	• Centroid as the cluster center is built in the middle of cluster member • At first, centroid is placed randomly then the iteration process used to improve the centroid location • Iteration also performed when a new data is placed into a data matrix • Cluster border is built in the middle of distance between cluster	**K-MEANS [22]** • Hard clustering • User need to define the value of K which is the number of clusters • The simplest algorithm but the K value need to set accurately **Fuzzy C Means, FCM [23]** • Soft Clustering where it assigns the cluster membership using degree of membership • It is a complex and resource consumption algorithm • Since it allows a data belong to more than one cluster, it faces some difficulty for a data that hold too much attribute

4. Comparative Analysis

CBIR need to be equipped with dataset as referencing material for CBIR to recognized a queried image and afterward placing the data into a relative cluster. The ability of CBIR to recognize an image and categorizing a dataset is rely on clustering techniques implemented. Designing the dataset will affect the data structure presentation. Some of the data need to be clustered softly as in **Table 3**.

Table 3. Hard and soft clustered data presentation.

Categorization style	Hard	Soft
Grouped data		
Tree structured (Subcategories)		

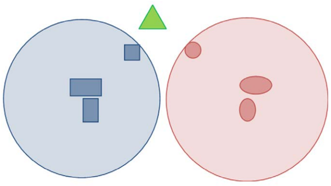
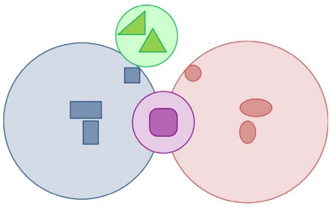
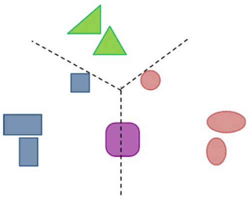
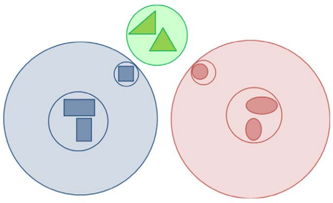
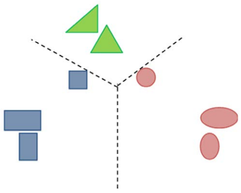
If you follow the "checklist" your paper will conform to the requirements of the publisher and facilitate a problem-free publication process.

Dataset matrix presentation that used by CBIR is not arrange as clean as illustrated in **Table 3** before, it need to be clustered in order to assign a data cluster membership. There is no ultimate complete clustering algorithm that suited for any dataset structure and requirement. Thus, the clustering algorithm must be selected properly for the best suit CBIR and datasets design. **Table 4** is the comparison of some clustering algorithms that suited for a specific dataset structure.

When a clustering algorithm wrongly implemented, it would result the output that does not meet the requirement and expectation. Well, K-Means is the simplest algorithm [24] that could execute clustering process faster than the other algorithm with minimal resources. For a CBIR that have unknown number of categories, the K-Means algorithm adaption will cause a data wrongly categorized. On the other hand, when a CBIR is required to employ a strict number of categories and cluster membership rule, others than K-Means algorithm type implementation might results a below expectation clustering.

Although clustering algorithm was developed long before the term big data being popular (around 90s), there are lots more research on data clustering research is emerge. From time to time, CBIR require being equipped with more sophisticated clustering algorithm due to the increasing requirement of image categorization alongside with CBIR application. CBIR developer need to choose the clustering algorithm wisely and comparative table as presented in **Table 4** should be a help.

Table 4. 2 Dimensional matrix presentations of dataset's designs.

Clustering algorithms	Dataset structure	Description
DBSCAN [19]		• Grouped Hard Clustering • No overlapping data attributes • Micro-cluster is considered as a different group is minimum density value reached • A data that not reached the minimum density, will treat as noise, a triangle also a shape, but it considered as noise
Expectation Maximization, EM [20]		• Tree Structured Soft Clustering • A data could be a member of many clusters • Micro-cluster is considered as a different group or to be placed in a greater group
Fuzzy C Means, FCM [23]		• Grouped Soft Clustering • A data could be member of many clusters based on degree of membership, normally the border point data
HDBSCAN [21]		• Tree Structured Hard Clustering • No overlapping data attributes • Subclustered is not an overlapping cluster, it just placed under a greater cluster
K-MEANS [22]		• Grouped Hard Clustering • No overlapping data attributes • No data will be treated as outliers • Must be have an exact number of clusters need to be form, otherwise the triangle will be the member of rectangle or round

5. Conclusion

In a CBIR development cycle, the purpose of a CBIR implementation needs to be decided first that would determine the dataset design and succeeding clustering algorithm to be adapted. Since there is no ultimate choice of a clustering algorithm for all clustering purpose, it is depending on CBIR developer wisdom to select and adapting the perfectly fit clustering algorithm. We hope as in this study analysis and findings become an aided reference on choosing the perfect clustering algorithm for CBIR development.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Ordonez, V., Deng, J., Choi, Y., Berg, A.C. and Berg, T.L. (2013) From Large Scale Image Categorization to Entry-Level Categories. *Proceedings of the IEEE International Conference on Computer Vision*.

[2] Cardoso, D.N.M., Muller, D.J., Alexandre, F., Neves, L.A.P., Trevisani, P.M.G. and Giraldi, G.A. (2013) Iterative Technique for Content-Based Image Retrieval Using Multiple SVM Ensembles. *J Clerk Maxwell, A Treatise on Electricity and Magnetism*, **2**, 68-73.

[3] Zloof, M.M. (1977) Query by Example: A Database Language. *IBM Systems Journal*, **16**, 324-343.

[4] Zhou, B., Khosla, A., Lapedriza, A., Torralba, A. and Oliva, A. (2017) Places: An Image Database for Deep Scene Understanding. *Journal of Vision*, **17**, 296.

[5] Bora, D.J., Gupta, D. and Kumar, A. (2014) A Comparative Study between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. arXiv: 1404.6059.

[6] Lv, Y.H., Ma, T., Tang, M., Cao, J., Tian, Y., Al-Dhelaan, A. and Al-Rodhaan, M. (2016) An Efficient and Scalable Density-Based Clustering Algorithm for Datasets with Complex Structures. *Neurocomputing*, **171**, 9-22.

[7] (2016) http://image-net.org/index

[8] Google, "Google Image" (2018) http://image.google.com

[9] Freitas, F.D.A., Peres, S.M., Lima, C.A.D.M. and Barbosa, F.V. (2014) Grammatical Facial Expressions Recognition with Machine Learning. *FLAIRS Conference*.

[10] Rothe, R., Timofte, R. and Gool, L.V. (2015) Dex: Deep Expectation of Apparent Age from a Single Image. *Proceedings of the IEEE International Conference on Computer Vision Workshops*.

[11] Heilbron, F.C., Escorcia, V., Ghanem, B. and Niebles, J.C. (2015) Activitynet: A Large-Scale Video Benchmark for Human Activity Understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[12] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet Classification with Deep Convolutional Neural Networks. *In Advances in Neural Information Processing Systems*.

[13] Xiao, H., Rasul, K. and Vollgraf, R. (2017) Fashion-Mnist: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv: 1708.07747.

[14] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*.

[15] Kussul, E. and Baidyk, T. (2004) Improved Method of Handwritten Digit Recognition Tested on MNIST Database. *Image and Vision Computing*, **22**, 971-981.

[16] Heitz, G., Elidan, G., Packer, B. and Koller, D. (2009) Shape-Based Object Localization for Descriptive Classification. *In Advances in Neural Information Processing Systems*.

[17] Digital Globe Inc. (2016) http://explore.digitalglobe.com/spacenet

[18] Fahad, A., AlShatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., Foufou, S. and Bouras, A. (2014) A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Transactions on Emerging Topics in Computing*, **2**, 267-279.

[19] Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD*, **96**, 226-231.

[20] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, *Se-*

ries B (*Methodological*), 1-38.

[21] Campello, R.J.G.B., Moulavi, D. and Sander, J. (2013) Density-Based Clustering Based on Hierarchical Density Estimates. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Heidelberg, Berlin.

[22] Macqueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*.

[23] Dunn, J.C. (1973) A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters.

[24] Rajalakshmi, K., Dhenakaran, D.S. and Roobin, N. (2015) Comparative Analysis of K-Means Algorithm in Disease Prediction. *International Journal of Science*, *Engineering and Technology Research* (*IJSETR*), **4**, 1-3.