sciendo

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **23**, No 2
Sofia • 2023 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2023-0010

Image Clustering and Feature Extraction by Utilizing an Improvised Unsupervised Learning Approach

R. Bhuvanya, M. Kavitha

VelTech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India
E-mails: bhuvanyaraghunathan@gmail.com kavitha@veltech.edu.in

***Abstract*:** *The need for information is gradually shifting from text to images due to the technology's growth and increase in digital images. It is quite challenging for people to find similar color images. To obtain similarity matching, the color of the image needs to be identified. This paper aims at various clustering techniques to identify the color of the digital image. Though many clustering techniques exist, this paper focuses on Fuzzy c-Means, Mean-Shift, and a hybrid technique that amalgamates the agglomerative hierarchies and k-Means, known as hKmeans to cluster the intensity of the image. Applying evaluation metrics of Mean Squared Error, Root Mean Squared Error, Mean Absolute Error, Homogeneity, Completeness, V-Score, and Peak signal-to-noise ratio it is proven that the results obtained demonstrate the good performance of the proposed technique. Then the color histogram is applied to identify the color and differentiate the color distribution on the original and clustered image.*

***Keywords*:** *Color histogram; evaluation metrics; feature extraction; image clustering; hybrid clustering.*

1. Introduction

As today's web is flooded with large sources of data, the World Wide Web also contains a massive amount of data which includes text, images, etc. To analyze the image data and extract the features, various image processing techniques, and data mining techniques play a predominant role. In addition to the above techniques, some other software tools such as R, SciKit, and Konstanz Information Miner (KNIME) can be used. Though there are various tools available there is no efficient tool separately available for image and audio [1, 2]. When analyzing the image data, it is necessary to apply the partitioning technique to divide the given data. As the image comes under the unlabeled category, clustering is a powerful tool that helps to accomplish the data partitioning task. Clustering also known as cluster analysis falls under the category of unsupervised Machine Learning (ML) tasks. The classification technique comes under supervised learning whereas some clustering techniques fall under both supervised and unsupervised learning. The goal of classification is

predictive and the goal of clustering is descriptive [3]. Clustering techniques can be applied when the instances are to be divided into natural groups. By applying the clustering technique, it is possible to detect and remove the noise. Also, it is extremely helpful to reduce the significant data from the dataset and this data reduction plays a vital role in image segmentation. The clustering methods are mainly classified into hard clustering and soft clustering. In hard clustering, while grouping the data items each item is assigned to only one cluster whereas soft clustering assigns the likelihood of datapoint to be in each of the clusters.

Unsupervised machine learning and data science depend heavily on cluster analysis. It is particularly useful for data mining and big data because, unlike supervised machine learning, it automatically identifies patterns in data without the need for labels. Clustering is the process of dividing the data points into a group such that similar data points are categorized into a cluster. It takes a set of items and focuses on their similarities and differences. The process of clustering can be described as follows. The initial stage is a pre-processing one, where the missing values can be handled followed by feature extraction [4], feature selection, and normalization. The next stage is proximity calculation, which measures the similarity and dissimilarity. The proximity can be obtained either directly or indirectly and it could be applied to nominal attributes, binary attributes, ordinal attributes, and numerical and mixed attributes. Finally, the applied clustering algorithm will categorize similar data points into a set of groups.

1.1. Outline of the proposed work

The proposed work focuses on color identification through clustering which helps to classify the color of the product by analyzing the intensity of cluster regions. The implemented work extracts images from the Flipkart dataset. Fig. 1 shows the outline of the proposed work. This paper focuses on Fuzzy C Means, Mean Shift, and a hybrid technique that combines agglomerative hierarchy and K-Means which are applied to color images that will categorize the pixel into groups. The clustering technique receives image data as input and generates clusters of pixels as output. To identify the color of the product various techniques such as color moments, color histograms, and color coherence vectors can be used. This paper focuses on identifying the color of the product by applying a color histogram on both the original and clustered image, which will represent the distribution of the color in an image.

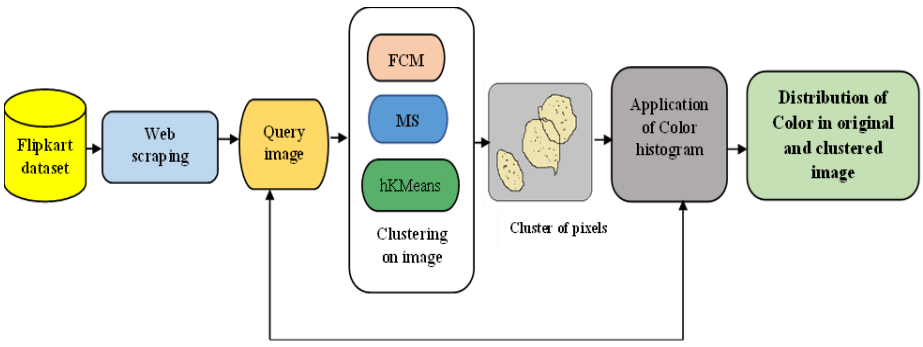


Fig. 1. Outline of the proposed work

1.2. Organization of the paper

The rest of the paper is organized as follows. Section 2 narrates the background study of clustering algorithms. Section 3 describes the clustering techniques which are applied to categorize the pixels and the color histogram technique to identify the distribution of color. Finally, Section 4 encompasses the efficiency analysis followed by the conclusion.

2. Background study

This section describes popular clustering techniques such as partitional, hierarchical, density-based, fuzzy, subspace, and model-based clustering. Cluster analysis differs from discriminant analysis as the former groups similar objects and the latter assign the objects to a group [5]. According to partitional clustering algorithms [6, 7], it decomposes the dataset into disjoint clusters. If the dataset has N points, then the partitioning methods construct K partitions of the data and ($K \leq N$). By minimizing the objective function, partitioned clustering algorithms divide the data into k clusters in an iterative manner. In the partitioning technique, two points are chosen to form two clusters; these two points are referred to as seed points. The selection of seed points is crucial. To form the right clusters, it is good to choose the points which are far away from each other. The important algorithms under partition clustering are k-Means, Partitioning Around Medoids (PAM), Clustering LARge Applications (CLARA), and Clustering Large Applications based on RANdomized Search (CLARANS) [8]. The major advantages of partition clustering are, that it is easy to implement and can work on large datasets and it is easy to infer the output. Also, partitional clustering can do the computation faster than hierarchical clustering and it produces tight clusters. Some advantages with partitional clustering exist, k-Means can handle only numerical data and the number of clusters (k-Value) must be defined at the outset by the user. It is difficult to predict the number of clusters and quality of clusters when using partitional clustering, and it cannot work with complex datasets.

The hierarchical clustering algorithms divide the dataset into a tree-like structure also called a dendrogram. The hierarchical clustering algorithm is further divided into Agglomerative and Divisive clustering. The agglomerative algorithm also known as AGglomerative NESting (AGNES) follows the bottom-up merge approach and the DIvisive Algorithm (DIANA) uses the top-down split approach [9]. The underlying principle of hierarchical clustering is that the solution is arranged in a hierarchy from n clusters to 1 cluster and vice versa. Here each data point is initially treated as an independent cluster, and from there, the algorithm iteratively aggregates the nearest clusters until a stopping requirement is met. Algorithms under hierarchical clustering are Balanced Iterative Reducing and Clustering using Hierarchies [10] (BIRCH), Clustering Using Representatives (CURE) [11], RObust Clustering using linKs (ROCK) [12] and Chameleon [13]. The advantage of hierarchical clustering is, the number of clusters (k) need not be specified prior. It is easy to implement and gives the best result in some cases. Though it can handle noise effectively, the computational cost is high when dealing with large datasets. It is difficult to identify the number of clusters through a dendrogram.

To deal with the arbitrary shape of clusters, density-based clustering is useful. This type of clustering algorithm's fundamental principle is that the high-density data points will form a separate cluster and the lower-dense regions will be categorized into different clusters [14]. The algorithm covers under density-based clustering are Density-Based Spatial Clustering (DBSCAN) [15], DENCAST [16], and Mean Shift [17]. Among these, DBSCAN is the popular density-based clustering algorithm. These algorithms are mainly applied to form clusters in nature-oriented applications such as spatial data and rivers as it is efficient in handling noise and outliers. It is not necessary to specify the number of clusters prior but it cannot work with large datasets due to the curse of dimensionality.

Fuzzy clustering is a type of soft clustering in which each data point may belong to one or more clusters based on the degree of membership. A data point's membership in a cluster increases with proximity to the cluster. Fuzzy cluster differs from k-Means clustering, as in k-Means the data points belong to a single cluster. The most popular algorithms under fuzzy clustering are Fuzzy c-Means [18] and Possibilistic c-Means [19]. This paper uses Fuzzy C-Means (FCM) to cluster the intensity of pixels in image data as FCM works efficiently with highly overlapped data and is especially applicable to image segmentation and bioinformatics. As fuzzy clustering belongs to centroid clustering, needs to specify the number of centroids prior and it is sensitive to noise and outliers.

To locate clusters in several subspaces within a dataset, subspace clustering is used [20, 21]. Subspace clustering is effective when used with high-dimensional datasets. There are two main branches of subspace clustering. Top-down algorithms locate an initial clustering in the complete set of dimensions and assess the subspace of each cluster progressively. In the bottom-up approach, dense regions are located in low-dimensional spaces, and these regions are then combined to form clusters. Algorithms under subspace clustering are CLIQUE (CLustering In QUEst) [22], and STING (STatistical INformation Grid) [23]. Subspace clustering is particularly effective in high-dimensional datasets. It has found numerous applications in image segmentation, motion segmentation, and face clustering. The density threshold is the same for both low and high-dimensional data, hence the clustering results may not be accurate in all cases.

The model-based clustering method makes an effort to improve how well the data and certain mathematical models fit together. Model-based clustering can automatically determine the ideal number of clusters. Hierarchical, Density-based clustering clusters the data points by computing either proximity or composition, whereas the data points are clustered using a distribution model-based clustering algorithm based on their likelihood of belonging to the same probability distribution. Expectation Maximization (EM) [24], conceptual clustering under machine learning [25], and Neural Networks Approach fall under this approach. As it uses a machine learning approach, the number of clusters does not need to be defined ahead of time and it works efficiently on real-time data.

The existing works of clustering algorithms [26, 27] are based on surveys by looking at the criteria such as the merits, knowledge about the domain, size of the dataset, number of cluster formations, and time complexity, and the above works

explain the application of clustering techniques in various image areas such as 3D imaging, medical, oceanography, remote sensing. The personalized recommendation in e-Commerce can be implemented [28] by learning the clustering representation. Also, the existing work [29], proposes High-Order Possibilistic c-Means (HOPCM) to cluster incomplete multimedia data.

3. Proposed work

As the process focuses on clustering the pixels on color images, the web scraping technique also known as web harvesting is applied here to extract the images from the product Uniform Resource Locator (URL). The dataset contains 20,000 records and each product may contain a maximum of five images, hence all the images need to be retrieved from the URL. The application of web scraping in the dataset extracts and downloads the images from the product URL. Fig. 2 shows the sample images collected from the product URL [30].



Fig. 2. Sample images collected from product URL

3.1. Fuzzy c-Means

It inherits the principles of fuzzy logic assigning each point a membership for each cluster center. FCM has some commonalities with the k-Means algorithm and the difference is, that the objective function of FCM allows different cluster membership and k-Means allows single cluster membership. FCM clustering is a type of soft clustering method. It will assign the membership to all data points by calculating the distance from the cluster center. The data points, in this case, may belong to multiple clusters, depending on the distance between the data object and the cluster centers. FCM algorithm works iteratively and partitions the image data into k partitions. FCM is based on the minimization of the objective function, which is described in the equation

$$J(U,V) = \sum_{i=1}^{n}\sum_{j=1}^{c}\left(\mu_{ij}\right)^{r}\left\|x_i - y_j\right\|^2, \quad (1)$$

where: μ_{ij} is the fuzzy partition matrix; $\|x_i - y_j\|$ is the Euclidean distance between i-th data and j-th cluster center; n is the number of data points; $c \in [2, n]$ is the number of clusters. J is the objective function in which V is the number of cluster centers. And $U = u_{ij} \in [0, 1]$, $i = 1, ..., n$, $j = 1, \ldots, c$, where each element u_{ij}, tells the degree to which element, x_i belongs to j-th cluster, and r is the fuzziness index, $r \in [1, \infty]$. An image with N pixels to be segmented into C clusters is denoted by, $A = \{A_1, A_2, \ldots, A_n\}$, where A_i, $i = 1, 2, \ldots, n$, represents the feature vector. For all $A_i \in A$, there is a cluster that belongs to C, such that $C = \{C_1, C_2, C_3, \ldots, C_n\}$.

Procedure FCM

Input: Image data A {A_1, A_2, …, A_n}, No of clusters (K)

Output: Cluster Centroid C={C_1,C_2, …, C_n}

Initialize Cluster centroid (c), Fuzziness degree (m), Termination parameters, membership_matrix Arr[]

Loop: for each i from 1 to k:

Loop: For each j from 1 to c:

Membershipmatrix+= $\mu_j * y_i$ *// Assign Membership Matrix*

End for

End for

Label: FCM

Loop: for every j, i from 1 to k:

Arr[i]= $(\mu_j * y_i)^m * y_i)$

Centroid$_j$ = Arr[i]/$(\mu_j * y_i)^m$ *// Computation of centroid*

// Dissimilarity calculation between data points and image centroid

Distance$_i$=sqrt($(x_2 - x_1)^2 + (y_2 - y_1)^2$)

//Update new membership matrix

for k from 1 to c:

Tmp+= $\left[\frac{1}{d_{k_i}}\right]^{1/(m-1)}$、

End for

for i, j from 1 to c:

$\mu_j(y_i)=[1/d_{ji}]^{1/(m-1)}$/Tmp

End for

End for

if the membership function is not constant goto FCM

End

3.2. Mean shift

Mean Shift (MS) is one of the popular clustering algorithms under unsupervised learning. The number of clusters need not be specified in advance. This algorithm performs well in image segmentation and video tracking [31]. It distributes the data points to the cluster by shifting towards the high-density region. In each iteration, the algorithm will assign the data points closest to the cluster center. Hence when the algorithm stops, each point is assigned to a cluster. To distribute the data points, Kernel Density Estimation (KDE) is used. Mathematically, a kernel is a weighting function that will apply weights for individual data points.

Mean shift is also known as a hill-climbing algorithm hence it needs to determine the direction where a sample must move to. Practically, it is computationally expensive to calculate for all the samples. Hence the parameter "bandwidth" helps to define an area around the samples for the mean shift to determine the most probable path. The term "quantile" (q) plays a major role in

estimating the bandwidth which defines a particular part of a dataset and it can cover the range of [0, 1]. The algorithm of MS clustering can be summarized as follows. Let $K(X_j - X_i)$ be considered as the kernel function which determines the weight of points to calculate the mean value and $N(X_i)$ be the neighborhood of X.

Procedure MS
Input: Image data A $\{A_1, A_2, \ldots, A_n\}$
Output: Clusters C $\{C_1, C_2, \ldots, C_n\}$
Read quantile, n_samples
X[n_samples] [n_features]
Label: MS
Loop: for i, j from 1 to n:
CALCULATE Tmp=$N(X_i)$*$K(X_j - X_i)$*X_j
$m(x_i)$=Tmp/ $N(X_i)$*$K(X_j - X_i)$ *// Calculation of Mean Shift*
Loop: for every data point x:
UPDATE $X \leftarrow m(x)$
MS=$m(x_i)$ – X *// Translation of density estimation*
End for
End for
Goto MS until convergence (or) all the points moved to higher density region.

3.3. Combining hierarchical clustering and k-Means

The proposed algorithm combines the technique of Hybrid agglomerative with k-Means. An arbitrary set of observations is selected as the initial centers in the k-Means method. This initial random selection of cluster centers has a significant impact on the final k-Means clustering solution. Hence at each computation, the result could be slightly different. Using a hybrid strategy that combines the k-Means and hierarchical clustering approaches can help prevent this. As this procedure integrates the agglomerative hierarchy and k-Means, it is also known as hKmeans.

3.3.1. k-Means

The most widely used partitioning algorithm is k-Means clustering, which clusters similar data and is commonly used in data mining [32]. KMeans is a hard and fast clustering algorithm, hence every data point here is not present in multiple clusters. It generates k clusters where k is the user-defined number. It is an iterative algorithm and it performs two major tasks namely Expectation-Maximization. The former will assign each data point to its closest centroid and the latter calculates the mean of all the points and reassign the new centroid [33]. The KMeans Algorithm is summarized below. Let X be the membership probability.

Procedure KMeans
Input: Image data A $\{A_1, A_2, \ldots, A_n\}$, No_of clusters ($K$)
Output: C=$\{C_1, C_2, C_3, \ldots, C_n\}$ set of centroids
Initialize C_j, j=1, 2, ..., n, *// Arbitrarily Initialize Random Centroid*
Label: KM

Loop: for i, j from 1 to n:

Expectation: A_i= argmin j $||X_i - C_j||^2$ *// Assign each point to its closest center*

Maximization: $C_j = \frac{1}{|A_i|} \Sigma_{x_i \in A_i} X_i$ *// Compute new cluster centers (mean)*

End for

If the convergence is not satisfied, goto KM

Return C_j

3.3.2. Agglomerative hierarchical algorithm

It is also known as AGglomerative NESting (AGNES). The process of clustering can be performed by combining similar patterns in the cluster set to generate a larger one. The results of the hierarchical algorithm can be represented in terms of the dendrogram. A dendrogram is a diagram that shows the hierarchical relationship between the data points which is organized in a tree structure. The leaf node of a tree represents the sub-cluster instead of a single data point. Hierarchical methods establish clustering in both top-down and bottom-up approaches. Clustering can be performed by calculating the distance measures which define how the similarity calculation of data points. Various distance measures such as Euclidean, Manhattan, Minkowski, and correlation can be implemented, where the default one is Euclidean.

The various formula to calculate the distance measures are given in Table 1. The appropriate choice of distance measure will determine the cluster shapes. After the distance measure is calculated, it is necessary to choose the cluster proximity. There are three measures of cluster proximity available namely simple, complete, and average [34]. The formula for measuring the cluster proximity is mentioned in Table 2.

Table 1. Types of distance measure and their description, where x_i and y_i specify the data points and p is the order parameter

No	Distance measure	Remarks	Formula
1	Euclidean	It can be calculated by taking the square distance between the points	$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
2	Manhattan	Manhattan distance measure can be implemented by considering the absolute difference	$\sum_{i=1}^{n} \lvert x_i - y_i \rvert$
3	Minkowski	The p-th root of the sum of the p-th powers of the differences of the components	$\left[\sum_{i=1}^{n} \lvert x_i - y_i \rvert^p\right]^{1/p}$

Table 2. Types of cluster proximity and its description, where x, and y refer to the two cluster points

No	Proximity measure	Remarks	Formula
1	Single linkage	Single linkage considers the distance between the clusters as the minimum distance	f=min($d(x, y)$)
2	Complete linkage	It considers the distance between the clusters as the minimum distance	f=max($d(x, y)$)
3	Average linkage	Here the proximity measure is defined by considering the distance between two clusters to be an average distance	f=average($d(x, y)$)

Procedure hierarchical_clustering

Input: {A_i :Image data}, i=1, 2, …, n), DIST (P, Q),
Output: C (Cluster hierarchy)
Initialize $X \leftarrow$ Null *// Initial cluster starts with an empty set*
Loop:
For each i from 1 to N: do
$X \leftarrow X \cup \{A_i\}$ *// Adding each data point to the initial cluster*
End for
$C \leftarrow X$
// Formation of Tree (C) by merging the clusters
Label: AGH
Loop: for every i:
// Distance calculation between two clusters
$P^*, Q^* \leftarrow$ argmin DIST (P, Q) *//P, Q refers two clusters*
// Delete them from the active cluster (X)
$X \leftarrow (X \setminus \{ P^*\} \setminus \{Q^*\})$
// compute the linkage (f) *and add them to Tree* (C)
$C \leftarrow X \cup \{P^* \cup Q^*\}$
If more than one cluster remains, goto AGH
End for
Return Tree C.

3.3.3. hKMeans

Procedure hKmeans

Input: Image data X_i {i=1, 2, …, n}, DIST (R, S)
Output: Set of cluster centroids (C), set of cluster labels (R)
$D \leftarrow$ hierarchical_clustering (V, E, d)
where V is the weighted graph, E is the set of edges and d is the distance
Initialize
Max_dist=NULL, $i \leftarrow 0$ and n[] where n is the number of nodes
Begin
for $i \leftarrow 2$ to n.length
dist$\leftarrow$get_distance(n_i, n_{i-1})
if (dist>max_dist) then
dist$\leftarrow$max_dist
$l \leftarrow$ level of node i in D
End
Assign $K \leftarrow$l+1
(C, R) $\leftarrow$ KMeans(V, K)

For unsupervised datasets, two important analytical methods include hierarchical and k-Means clustering. Hence, in a nutshell, the combined technique uses hierarchical clustering to cluster around half the data, followed by k-Means for the remaining half in a single cycle. In the beginning, the image that has been taken from the dataset will be fed into a hierarchical method, where the background will be

suppressed and the foreground region will be given to the k-Means clustering algorithm. k-Means can offer effective clustering results because it only considers the foreground area. This combined strategy offers two advantages where the first one is, it is not necessary to assign the arbitrary value of *K*, and also the initial centroids will be created in a much better way.

As the work focuses on color identification, if the data points are clustered, color identification will be much simpler. Hence the proposed work focuses on Fuzzy c-Means, and Mean Shift which play a dominant role in image processing. In addition to the above, a new algorithm is proposed "hkmeans" (combination of hierarchical clustering with k-Means) which performs significantly better when compared with the existing works of Fuzzy c-Means and Mean shift. The results of all the clustering algorithms are depicted in the upcoming section.

3.4. Feature extraction

The color feature is a popular visual characteristic in a digital image. Utilizing the color capability allows one to portray the image's visual content by obtaining the color information. The method most frequently used for color extraction is "color histogram". This paper focuses on the color histogram technique to extract color information from the images.

Color histogram is a visual depiction of how colors are distributed within an image. By counting the occurrence of each possible color of the associated color model inside the image, the information in a histogram is generated. Red, blue, and green values are the three important values that make up each pixel in a digital image and determine its color. These numbers can have values between 0 and 255, with 0 denoting none and 255 denoting the highest value. The number of pixels representing each color in the image is determined using a histogram. In other words, the intensity distribution of an image is displayed via the histogram plot.

4. Experiments and evaluation

4.1. Datasets

The dataset used to evaluate is a pre-crawled dataset that has been created by extracting data from Flipkart.com, a leading e-Commerce site. This dataset has the fields such as product_url, product name, product category tree, pid, retail-price, discounted price, image, is_FK_Advantage_product, description of the product, product rating, overall rating, brand, and product specification. This dataset has 20,000 records of product details. Fig. 3 shows the label distribution of the Flipkart dataset [30].

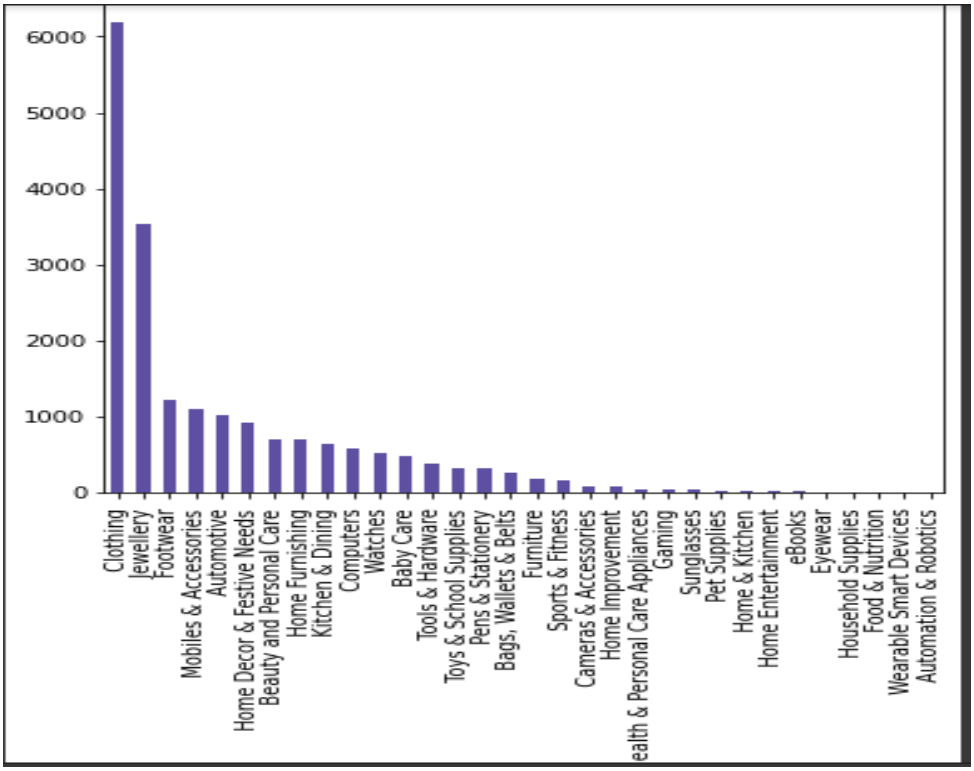


Fig. 3. Label distribution of the dataset

4.2. Performance analysis of clustering techniques

This section encompasses the results of all the clustering algorithms and the evaluation metrics applied. From the above label distribution, various categories of products are chosen and applied to the clustering algorithms. Table 3 depicts the query image and the results of all the clustering techniques such as Fuzzy-C-Means, Mean shift, and hKmeans, respectively.

From the above results, visually it is shown that the hKmeans technique produces better results. To prove the results mathematically, various evaluation metrics have been applied such as Mean Absolute Error (MAE), Mean Square Error (MSE), Root Mean Square Error (RMSE), Homogeneity, Completeness, V-Score, Peak Signal to Noise Ratio (PSNR) to compare the performance of the applied techniques. The formulae and the descriptions* of all the evaluation metrics are represented in Table 4. The term P and R represents the predicted and actual value respectively.

The following Table 5 exemplifies the detailed comparative results of all the implemented algorithms mathematically. For instance, the images obtained from original datasets are numbered 1 to 7 accordingly.

Table 3. Results of the clustering techniques

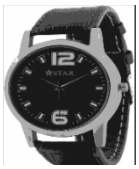
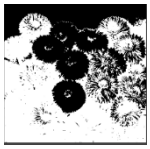
Description	Category	Query image	Results of Fuzzy c-Means	Results of Mean shift	Results of hKmeans
Image 1	Clothing				
Image 2	Home furnishing				
Image 3	Watches				
Image 4	Sports & fitness				
Image 5	Jewellery				
Image 6	Footwear				
Image 7	Home décor & festive needs				

Table 4. Evaluation metrics applied

No	Evaluation Metric	Remarks	Equation
1	Mean Absolute Error (MAE)	Absolute Accuracy Error is the amount of error in measurement. It is the difference between the predicted and the actual rating whereas the MAE is the average of all absolute errors [35]	$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}\lvert P - R\rvert$
2	Mean Square Error (MSE)	MSE is calculated by taking the difference between the actual and predicted ratings and squaring them. The squaring is done to remove the negative signs. Also, MSE paves the way to understand and calculate RMSE	$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(P - R)^2$
3	Root Mean Square Error (RMSE)	The loss function is high, which is the main problem with the MSE. Therefore, the RMSE is designed to lower the loss function by subtracting the root value from the MSE's obtained value	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P - R)^2}$
4	Homogeneity	When all the cluster's data points are members of the same class, the clustering outcome is homogeneous	$h = 1 - \frac{H(Y_{\text{true}}\lvert Y_{\text{pred}})}{H(Y_{\text{true}})}$
5	Completeness	Completeness describes the closeness of the clustering algorithm to this perfection	$c = 1 - \frac{H(Y_{\text{true}}\lvert Y_{\text{pred}})}{H(Y_{\text{pred}})}$
6	V-Score	Explicitly assesses how well the homogeneity and completeness criteria have been met	$V =$ (1 + beta)*h*c/(beta*h+c)
7	Peak Signal to Noise Ratio (PSNR)	Peak Signal/Noise Ratio (PSNR) measures the quality reconstruction of the image. Hence the high value of PSNR is directly proportional to the high-quality image and it is usually expressed in log scale	PSNR= $20.\log_{10}$ (Max $_I$ / √MSE) where, Max $_I$ is the maximum pixel value of an image

From Table 5, it can be concluded that the proposed algorithm (hKmeans) produces higher-quality results for all the images compared with FCM and MS respectively. In addition to the above, to analyze the results of clustering, a scatter plot is used. It is one of the important graphical tools that play an important role in correlation and regression [36]. A few instances of data points are chosen from the clustered region, to represent the scatter plot. Fig. 4 represents the formation of cluster and scatter plots for all the applied above techniques.

Further, the feature extraction technique known as the color histogram is applied, which depicts the distribution of color on the images. Table 6 shows the color distribution of the query and clustered images respectively.

The color histogram highlights the color ratio and counts how many pixels are in each of the bins. The bin values are in the range of 0-255, which corresponds to the *X*-axis, and pixel counts are specified on *Y*-axis. Each bin in a histogram represents a certain range of intensity values. By looking at each pixel in the image and placing it in a bin based on its intensity, the histogram is calculated. From the above Table 6, it is evident that the number of pixels is reduced by applying the clustering technique and color identification can be done efficiently when the number of pixels is reduced.

Table 5. Comparative performance evaluation of FCM, MS, hKmeans

Image data	Algorithms	MAE	MSE	RMSE	Homogeneity	Completeness	V-Score	PSNR
Image 1	Fuzzy c-Means	218.97	112.93	10.62	0.2271	1.0	0.3702	27.60
	Mean Shift	218.37	112.97	10.62	0.2783	1.0	0.4066	27.60
	hKmeans	199.00	56.42	7.51	0.2700	1.0	0.4355	30.61
Image 2	Fuzzy c-Means	147.66	39.28	6.26	0.1944	1.0	0.325	32.18
	Mean Shift	147.53	36.97	6.08	0.3018	0.90	0.4530	32.45
	hKmeans	121.88	28.78	5.36	0.3119	0.99	0.475	33.53
Image 3	Fuzzy c-Means	149.29	42.31	6.50	0.1516	0.99	0.263	31.86
	Mean Shift	143.05	40.16	6.33	0.266	0.99	0.409	32.09
	hKmeans	116.25	31.32	5.59	0.316	1.0	0.481	33.17
Image 4	Fuzzy c-Means	188.02	44.11	6.64	0.205	1.0	0.340	31.68
	Mean Shift	185.56	44.51	6.67	0.319	1.0	0.484	31.64
	hKmeans	157.84	36.47	6.03	0.271	0.99	0.490	32.51
Image 5	Fuzzy c-Means	161.47	48.78	6.98	0.169	1.0	0.290	31.24
	Mean Shift	161.37	46.70	6.83	0.229	0.77	0.353	31.43
	hKmeans	142.34	32.90	5.73	0.335	0.99	0.50	32.95
Image 6	Fuzzy c-Means	174.71	46.77	6.83	0.188	0.88	0.313	30.43
	Mean Shift	170.53	45.43	6.74	0.277	1.0	0.434	31.16
	hKmeans	144.48	34.45	5.86	0.302	1.0	0.450	33.45
Image 7	Fuzzy c-Means	121.33	45.42	6.73	0.129	0.99	0.228	31.16
	Mean Shift	126.47	46.52	6.82	0.141	0.56	0.226	33.26
	hKmeans	110.94	35.46	5.95	0.189	0.99	0.318	35.55

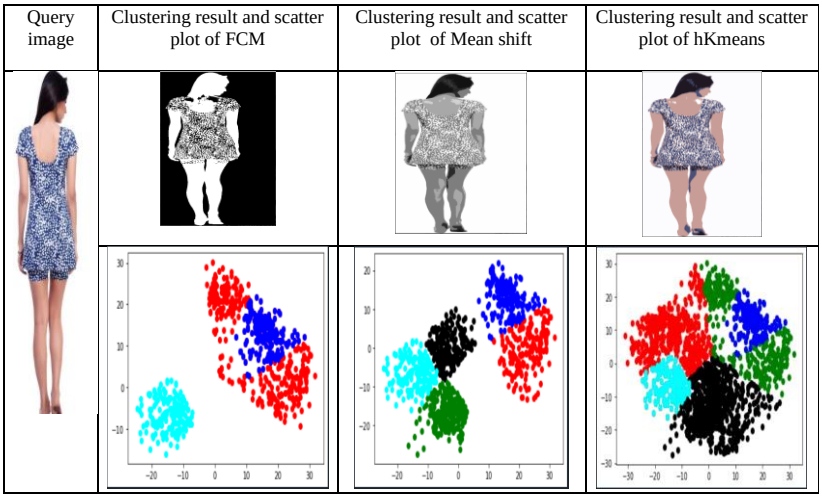
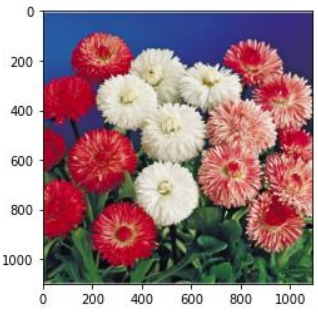
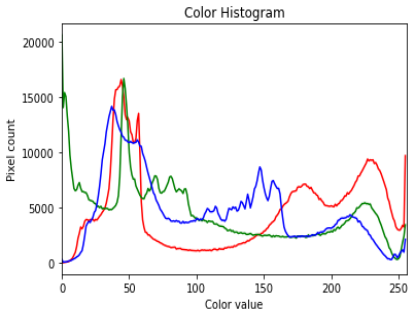
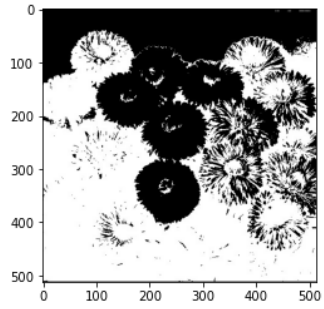
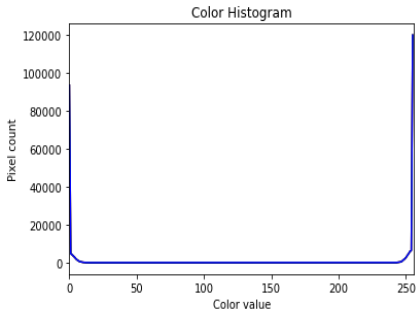
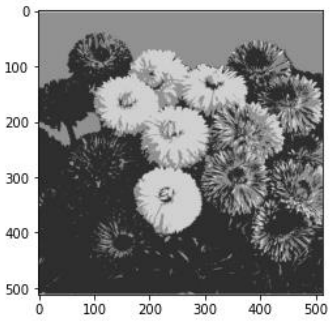
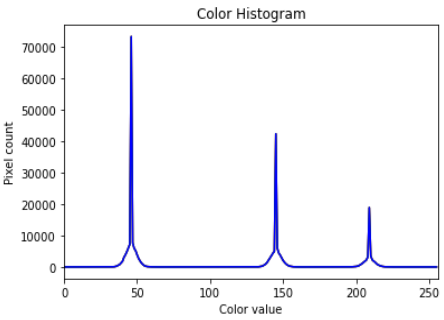
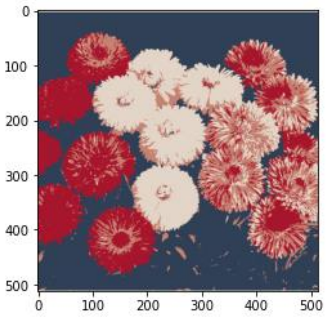
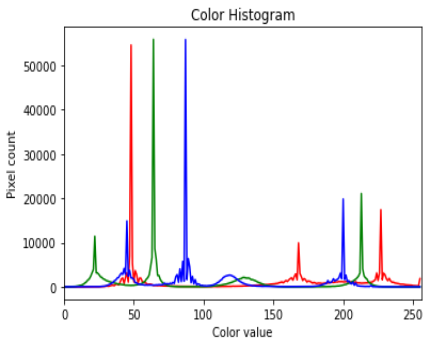


Fig. 4. Cluster regions identified by FCM, MS, and hKmeans

Table 6. Color histogram of query and clustered images

1	Query image		
2	Fuzzy-c-Means		
3	Mean Shift		
4	hKmeans		

5. Conclusion and future work

It is proven that the adopted work can be taken as an initial step for the color identification process and it could be applied in real-world e-Commerce applications to retrieve similar color products. The color histogram methodology applied here is considered as the hand-crafted feature extraction technique. Hence in the future, the work will be carried out by adopting deep learning techniques, which will pave the way for similarity-based image retrieval in real-time e-Commerce applications as deep learning can extract the shape, size, and texture in addition to the color of the product.

References

1. Jovic, A., K. Brkic, N. Bogunovic. An Overview of Free Software Tools for General Data Mining. – In: Proc. of 37th IEEE International Convention on Information and Communication Technology, Electronics, and Microelectronics, 2014, pp. 1112-1117.
2. Mikut, R., M. Reischl. Data Mining Tools. – Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. **1**, 2011, No 5, pp. 431-443.
3. Rokach, L. A Survey of Clustering Algorithms. – In: Data Mining and Knowledge Discovery Handbook. 2nd Edition. 2010, pp. 269-298.
4. Wazarkar, S., B. Keshavamurthy, A. Hussain. Probabilistic Classifier for Fashion Image Grouping Using Multilayer Feature Extraction Model. – International Journal of Web Services Research, Vol. **15**, 2017, pp. 89-104.
5. Kaufman, L., P. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. New York, John Wiley & Sons, 2009.
6. Jin, X., J. Han. Partitional Clustering. – In: Encyclopedia of Machine Learning. Boston, MA, Springer, 2011.
   https://doi.org/10.1007/978-0-387-30164-8_631.
7. Salem, S. B., S. Naouali, Z. Chtourou. A Fast and Effective Partitional Clustering Algorithm for Large Categorical Datasets Using a k-Means-Based Approach. – Computers & Electrical Engineering, Vol. **68**, 2018, pp. 463-483.
8. Schubert, E., P. J. Rousseeuw. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms, Similarity Search and Applications. – In: Lecture Notes in Computer Science, 2019, 11807.
9. Marjan, K. R., A. Zahra, E. Nasibeh. A Survey of Hierarchical Clustering Algorithms. – The Journal of Mathematics and Computer Science, Vol. **5**, 2012, pp. 229-240.
10. Madan, S. K., J. Dana. Modified Balanced Iterative Reducing and Clustering Using Hierarchies (m-BIRCH) for Visual Clustering. – Pattern Analysis and Applications, Vol. **19**, 2016, pp. 1023-1040.
11. Bouguettaya, A., Q. Yu, X. Liu, X. Zhou, A. Song. Efficient Agglomerative Hierarchical Clustering. – Expert Systems with Applications, Vol. **42**, 2015, No 5, pp. 2785-2797.
12. Guha, S., R. Rastogi, K. Shim. Rock: A Robust Clustering Algorithm for Categorical Attributes. – Information Systems, Vol. **25,** 2000, No 5, pp. 345-366.
13. Karypis, G., E.-H. Han, V. Kumar. Chameleon: Hierarchical Clustering Using Dynamic Modelling. – Computer, Vol. **32**, 1999, No 8, pp. 68-75.
14. Kriegel, H.-P., P. Kröger, J. Sander, A. Zimek. Density-Based Clustering. – Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. **1**, 2011, pp. 231-240.
15. Yu, H., L. Y. Chen, J. T. Yao, X. N. Wang. A Three-Way Clustering Method Based on an Improved DBSCAN Algorithm. – Physica A: Statistical Mechanics and Its Applications, Vol. **535**, 2019, 122289.
16. Al-Jabery, K. K., T. Obafemi-Ajayi, G. R. Olbricht, D. C. Wunsch II. Computational Learning Approaches to Data Analytics in Biomedical Applications. Elsevier, 2019.

17. G u o, Y., A. Ş e n g ü r, Y. A k b u l u t, A. S h i p l e y. An Effective Color Image Segmentation Approach Using Neutrosophic Adaptive Mean Shift Clustering. – Measurement, Vol. **119**, 2018, pp. 28-40.
18. B o r l e a, I.-D., R.-E. P r e c u p, A.-B. B o r l e a, D. I e r c a n. A Unified Form of Fuzzy c-Means and k-Means Algorithms and Its Partitional Implementation. – Knowledge-Based Systems, Vol. **214**, 2021, 106731.
19. A s k a r i, S. Fuzzy c-Means Clustering Algorithm for Data with Unequal Cluster Sizes and Contaminated with Noise and Outliers: Review and Development. – Expert Systems with Applications, Vol. **165**, 2021,113856.
20. K r i e g e l, H.-P., P. K r ö g e r, A. Z i m e k. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. – ACM Transactions on Knowledge Discovery from Data, Vol. **3**, 2009, No 1, pp. 1-58.
21. K r ö g e r a n d, P., A. Z i m e k. Subspace Clustering Techniques. – In: Encyclopedia of Database Systems, Boston, MA, Springer, 2009.
https://doi.org/10.1007/978-0-387-39940-9_607.
22. B a o, X., L. W a n g. A Clique-Based Approach for Co-Location Pattern Mining. – Information Sciences, Vol. **490**, 2019, pp. 244-264.
23. A g r a w a l, R., J. G e h r k e, D. G u n o p u l o s, P. R a g h a v a n. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. – In: Proc. of ACM Sigmod International Conference on Management of Data, Vol. **27**, 1998, pp. 94-105
24. W u, C. J. On the Convergence Properties of the EM Algorithm. – In: The Annals of Statistics, 1983, pp. 95-103.
25. C h e n g, Y., K. S. F u. Conceptual Clustering in Knowledge Organization. – In: PAMI 7, 1998, pp. 592- 598.
26. H e, L., L.-d. W u, Y.-c. C a i. Survey of Clustering Algorithms in Data Mining. – Application Research of Computers, Vol. **1**, 2007, pp. 10-13.
27. S i n g h a l, G., S. P a n w a r, K. J a i n, D. B a n g a. A Comparative Study of Data Clustering Algorithms. – International Journal of Computer Applications, Vol. **83**, 2013, No 15, pp. 41-46.
28. W a n g, K., T. Z h a n g, T. X u e, Y. L u, S.-G. N a. e-Commerce Personalized Recommendation Analysis by Deeply-Learned Clustering. – Journal of Visual Communication and Image Representation, Vol. **71**, 2020, 102735.
29. Z h a n g, Q., L. T. Y a n g, Z. C h e n, F. X i a. A High-Order Possibilistic c-Means Algorithm for Clustering Incomplete Multimedia Data. – IEEE Systems Journal, Vol. **11**, 2017, No 4, pp. 2160-2169.
30. **https://www.kaggle.com/PromptCloudHQ/flipkart-products**
31. L i u, Y., S. Z. L i, W. W u, R. H u a n g. Dynamics of a Mean-Shift-Like Algorithm and Its Applications on Clustering. – Information Processing Letters, Vol. **113**, 2013, No 1-2, pp. 8-16.
32. L o n g, Z.-Z., G. X u, J. D u, H. Z h u, T. Y a n, Y.-F. Y u. Flexible Subspace Clustering: A Joint Feature Selection and k-Means Clustering Framework. – Big Data Research, Vol. **23**, 2021, 100170.
33. Y a o, H., Q. D u a n, D. L i, J. W a n g. An Improved k-Means Clustering Algorithm for Fish Image Segmentation. – Mathematical and Computer Modelling, Vol. **58**, 2013, No 3-4, pp. 790-798.
34. G i l-G a r c i a, R. J., J. M. B a d i a-C o n t e l l e s, A. P o n s-P o r r a t a. A General Framework for Agglomerative Hierarchical Clustering Algorithms. – In: Proc. of 18th International Conference on Pattern Recognition,Vol. **2**, 2006, pp. 569-572.
35. H e r l o c k e r, J., J. K o n s t a n, L. T e r v e e n, J. C. L u i, T. R i e d l. Evaluating Collaborative Filtering Recommender Systems. – ACM Transactions on Information Systems, Vol. **22**, 2004, pp. 5-53.
36. S a i n a n i, K. L. The Value of Scatter Plots. – PM&R, Vol. **8**, 2016, No 12, pp. 1213-1217.

Received: 05.10.2022; Second Version: 22.03.2023; Accepted: 14.04.2023