Contribution Based Clustering Algorithm for Content-Based Image Retrieval (CBIR)

I. Srinivas Rao, A. Gauthami Latha
1,2Dept of CSE, VIIT Duvvada, India

Abstract
Clustering is a form of unsupervised classification that aims at grouping data points based on similarity. In this paper, we propose a new partitional clustering algorithm based on the notion of ‘contribution of a data point’. We apply the algorithm to content-based image retrieval and compare its performance with that of the k-means clustering algorithm. Unlike the k-means algorithm, our algorithm optimizes on both intra-cluster and inter-cluster similarity measures. It has three passes and each pass has the same time complexity as an iteration in the k-means algorithm. Content based image retrieval (CBIR) is done using the image feature set extracted from Haar Wavelets applied on the image at various levels of Decomposition. Here the database image features are extracted by applying Haar Wavelets on gray plane (average of red, green and blue) and color planes (red, green and blue components). Our experiments on a bench mark image data set reveal that our algorithm improves on the recall at the cost of precision. The results show that precision and recall of Haar Wavelets are better than complete Haar transform based CBIR, which proves that Haar Wavelets gives better discrimination capability in image retrieval at higher query execution speed, per higher level Haar Wavelets.

Keywords
Content-based Image Retrieval (CBIR); Clustering Contribution; K-Means Algorithm; Haar Wavelets

I. Introduction
A cluster is a collection of data points that are similar to one another within the same cluster and dissimilar to data points in other clusters [1]. Clustering is a method of unsupervised classification, where data points are grouped into clusters based on their similarity. The goal of a clustering algorithm is to maximize the intra-cluster similarity and minimize the inter-cluster similarity. Clustering algorithms can be broadly classified into five types: 1. Partitional clustering, 2. Hierarchical clustering, 3. Density-based clustering, 4. Grid-based clustering and 5. Model-based clustering [1]. Partitional and hierarchical clustering are the most widely used forms of clustering. In partition clustering, the set of n data points are partitioned into k non-empty clusters, where k ≤ n. In the case of hierarchical clustering, the data points are organized into a hierarchical structure, resulting in a binary tree or dendrogram [2]. In this paper, we propose a new clustering algorithm, which would come under the category of partitional clustering algorithms. Two commonly used methods for partitioning data points include the k-means method [3, 4] and the k-medoids method [5]. In the k-means method, each cluster is represented by its centroid or the mean of all data points in the cluster. In the case of the k-medoids method, each cluster is represented by a data point close to the centroid of the cluster. Apart from these methods, there has been lots of work on fuzzy partitioning methods [6] and partition methods for large scale datasets [7]. We use the notion of ‘contribution of a data point’ for partitional clustering. The resultant algorithm requires only three passes and we show that the time complexity of each pass is same as that of a single iteration of the k-means clustering algorithm. While the k-means algorithm optimizes only on the intra-cluster similarity, our algorithm also optimizes on the inter-cluster similarity.

Clustering has widespread applications in image processing. Color-based clustering techniques have proved useful in image segmentation [13]. The k-means algorithm is quite popular for this purpose. Clustering based on visual content of images is an area that has been extensively researched for several years [14]. This finds application in image retrieval. Content-Based Image Retrieval (CBIR) aims at finding images of interest from a large image database using the visual content of the images. Traditional approaches to image retrieval were text-based, where individual images had to be annotated with textual descriptions [8]. Since this is a tedious manual task, image retrieval based on visual content is very essential. Organizing the retrieved search results into clusters is an intuitive form of content representation [14] and facilitates user’s browsing of images [15]. Image clustering can also be used to optimize the performance of a CBIR system [16]. While the performance of a number of clustering algorithms in image retrieval have been analyzed in previous works [17-20], we apply our proposed algorithm to CBIR and compare its performance with that of the k-means clustering algorithm. The features of the query image are compared with the features of all images in the large database using various similarity measures.

A. Haar Wavelet
The Haar transform is derived from the Haar matrix. The Haar transform is separable and can be expressed in matrix form

\[ F = [H] [f] [H]^T \]

where F is an N x N image, H is an N x N Haar transform matrix and F is the resulting N x N transformed image. The transformation H contains the Haar basis function h_k(t) which are defined over the continuous closed interval \( t \in [0, 1] \). The Haar basis functions are When k=0, the Haar function is defined as a constant

\[ h_0(t) = 1/\sqrt{N} \]

When k>0, the Haar Function is defined as

\[ h_k(t) = \begin{cases} 
2^{-p/2} & \text{for } 0, \frac{1}{2}, \frac{1}{4}, \ldots, \frac{1}{2^p}, \frac{1}{2^p} q = \left\lfloor \frac{1}{2^p} q \right\rfloor \\
0 & \text{otherwise}
\end{cases} \]

(1)

Where 0 ≤ p ≤ \log_2 N and 1 ≤ q ≤ 2^p

The proposed algorithm is outlined in Section III and its time complexity is analyzed in Section IV. In Section V, we describe how clustering can be applied to content-based image retrieval and present our experiments and results in section VI. Section VII contains some conclusions along with future directions of research.

II. Contribution-Based Clustering
Partitional clustering aims at partitioning a group of data points into disjoint clusters optimizing a specific criterion [2]. When the number...
of data points is large, a brute force enumeration of all possible combinations would be computationally expensive. Instead, heuristic methods are applied to find the optimal partitioning. The most popular criterion function used for partitional clustering is the sum of squared error function given by:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} (x - m_i)^2$$

Where \( k \) is the number of clusters, \( C_i \) is the \( i \)-th cluster, \( x \) is a data point and \( m_i \) is the centroid of the \( i \)-th cluster.

A widely used squared-error based algorithm is the k-means clustering algorithm [2]. In this paper, we propose a clustering algorithm similar to the k-means algorithm. We define the contribution of a data point belonging to a cluster as the impact that it has on the quality of the cluster. This metric is then used to obtain an optimal set of \( 'k' \) clusters from the given set of data points.

The notion of contribution has its origin in game theory [9]. A recent work by Garg [10] focuses on the merger of game theory and data clustering. Garg mapped cluster formation to coalition formation in cooperative games and used the solution concept of Shapely value to find the optimal number of clusters for a given set of data points. While his work uses the concept of contribution to find the optimal cluster number, we use it in a different manner for optimal partitioning of the data points into a fixed number of clusters.

Given a cluster \( C_i \) with \( n \) points and centroids \( m_i \), the dispersion of \( C_i \) is given by

$$dispersion(C_i) = \frac{1}{n} \sum_{x \in C_i} (x - m_i)^2$$

The contribution of a point \( x \in C_i \) is measured as

$$Contribution(x, C_i) = dispersion(C_i - \{x\}) - dispersion(C_i)$$

Clearly, if the contribution of a data point is negative, it has an adverse impact on its cluster. On the other hand, a positive contribution indicates that the removal of the point from the cluster would degrade its quality.

In our work, we propose a clustering algorithm that treats points with negative contribution different from those with positive contribution. In the case of a negative contribution point, the point is shifted to a cluster, where its contribution is the highest, possibly positive. On the other hand, for a positive contribution point, a multi-objective optimization criterion is taken, where we try to optimize both the intra-cluster and inter-cluster dispersion measures.

### III. Algorithm

We now outline the proposed contribution-based clustering algorithm. It optimizes on two measures, namely the intra-cluster dispersion given by

$$\alpha = \frac{1}{n} \sum_{x \in C_i} (x - m_i)^2$$

and the inter-cluster dispersion given by

$$\beta = \frac{1}{k} \sum_{i=1}^{k} (m_i - \bar{m})^2$$

Where \( k \) is the number of clusters and is the mean of all Centroids. The algorithm tries to minimize \( \alpha \) and maximize \( \beta \). The three steps (passes) in the algorithm are described below.

1. **Initialization**
   
   Randomly select \( k \) centroids \( (m_1, m_2, ..., m_k) \)
   
   For each point \( x \)
   
   Find \( 1 \leq l \leq k \) such that distance\((x, m_l)\) is minimum
   
   Add \( x \) to cluster \( C_l \) and update centroid \( m_l \)
   
   End For

2. **Negative Contribution Points**

   For each cluster \( C_l \)
   
   For each point \( x \in C_l \)
   
   If \( Contribution(x, C_l) < 0 \)
   
   Move \( x \) to a cluster \( C_p \) such that \( Contribution(x, C_p) \) is maximum
   
   Update centroid \( m_p \)
   
   End If
   
   End For
   
   End For

3. **Positive Contribution Points**

   For each cluster \( C_l \)
   
   For each point \( x \in C_l \)
   
   If \( Contribution(x, C_l) \geq 0 \)
   
   Move \( x \) to a cluster \( C_p \) such that \( \alpha_{new} + \beta_{new} \) is maximum
   
   End if
   
   End for
   
   End for

Note that \( \alpha_{new} \) and \( \beta_{new} \) are values of \( \alpha \) and \( \beta \) after the point \( x \) is moved to cluster \( C_p \).

### IV. Complexity Analysis

Consider a cluster the \( C_i \) with \( n \) points. Let \( x_i \) denote the \( i \)-th point in the cluster and let \( d \) be the dimension of each point. Then its centroids is given by

$$m_i = \left( \frac{1}{n} \sum_{i=1}^{n} x_{ij} \right)_{j=1}^d$$

Let us consider a point \( x_{n+1} \) outside the cluster. The contribution of the point if it were added to the cluster would be contribution \( (x_{n+1}, C_i) = dispersion(C_i) - dispersion(C_i - \{x_{n+1}\}) \).

Note that

$$dispersion(C_i) = \frac{1}{n} \sum_{i=1}^{d} \left( x_{ij}^2 + \frac{1}{n} \left( \sum_{k=1}^{n} x_{kj} \right)^2 \right) - \frac{2}{n} x_{ij} \sum_{k=1}^{n} x_{kj}$$

Note that

$$\alpha_{new} = \alpha - \frac{\alpha_{old}}{\alpha + \beta \cdot \beta_{old}}$$

$$\beta_{new} = \beta - \frac{\beta_{old}}{\alpha + \beta \cdot \beta_{old}}$$

where \( \alpha \) and \( \beta \) are the intra-cluster and inter-cluster dispersion measures respectively.
Substituting, we get
\[
\sum_{i=1}^{n} \sum_{j=1}^{d} x_i^j = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} x_i^j + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{d} \left( \sum_{k=1}^{n} x_k^j \right) - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{d} x_i^j \sum_{k=1}^{n} x_k^j
\]

and
\[
\text{dispersion}(c_j) = \frac{1}{n+1} \sum_{i=1}^{n} \sum_{j=1}^{d} x_i^j - \frac{2}{(n+1)^2} \sum_{i=1}^{n} \sum_{j=1}^{d} \left( \sum_{k=1}^{n} x_k^j \right)^2 - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{d} x_i^j \sum_{k=1}^{n} x_k^j.
\]

Substituting, we get contribution
\[
(x_{w+1}, c_i) = \sum_{i=1}^{n} \sum_{j=1}^{d} x_i^j \left( \frac{1}{n} - \frac{1}{n+1} \right)
\]
\[
\frac{1}{n+1} \sum_{i=1}^{n} \sum_{j=1}^{d} x_i^j + \sum_{i=1}^{n} \sum_{j=1}^{d} \left( \sum_{k=1}^{n} x_k^j \right) \left( \frac{1}{n} - \frac{1}{n+1} \right) - \frac{1}{(n+1)^2} \sum_{j=1}^{d} \sum_{x_{w+1}^j} \left( x_{w+1,j}^j \right)^2
\]
\[
- \frac{1}{n} \sum_{i=1}^{n} x_i^j - \frac{1}{(n+1)^2} \sum_{i=1}^{n} x_i^j
\]
\[
= \sum_{i=1}^{n} x_i^j \left( \frac{1}{n} - \frac{1}{n+1} \right) - \frac{1}{n} \sum_{j=1}^{d} (c_j + x_{w+1,j})^2
\]
\[
- 2 \sum_{j=1}^{d} S_j \left( \frac{1}{n} - \frac{1}{n+1} \right) x_{w+1,j}^j.
\]

Where \( ss = \sum_{i=1}^{n} \sum_{j=1}^{d} x_i^j \) and for \( 1 \leq j \leq d, s_j = \sum_{i=1}^{n} x_{i,j} \).

It can be clearly seen that the contribution of a point in a cluster can be calculated in \( O(d) \) time complexity provided the values of \( SS \) and \( S_j \), \( 1 \leq j \leq d \), are recorded and updated after the transfer of each point. Similarly, it can be shown that the inter-cluster and intra-cluster dispersions can be calculated in \( O(d) \) time complexity. Since the algorithm spends most of its time in calculating contributions, the time complexity of each step is \( O(Nkd) \), where \( N \) is the total number of data points. This is same as the time complexity of a single iteration of the \( k \)-means algorithm. Note that the complexity becomes near linear when \( N \) is quite large compared to \( k \) and \( d \) [2]. The space complexity of the algorithm is \( O(N + k + d) \).

V. Content-Based Image Retrieval

Content-based image retrieval is a technique which uses visual content to search images from large-scale databases according to users’ interest [8]. A common method of querying a content-based image retrieval system is to provide an example image. The system then retrieves all images in the database that are similar in content with the query image.

VI. Experiments and Results

Our test data consisted of 777 images belonging to 18 categories obtained from the University of Washington’s Object and Concept Recognition for CBIR research project image dataset [21]. Each category contained varying number of images. All the images contained a textual description mentioning the salient foreground objects.

The images were clustered using our algorithm with the initial centroids chosen at random. The cluster whose centroids was closest in distance to the given test image was determined and the images belonging to the cluster were retrieved. The results were then compared with images retrieved using the \( k \)-means clustering algorithm with the same set of initial centroids. We performed two experiments to evaluate the effectiveness of our proposed approach over a large image database. A retrieved image is considered relevant if it belongs to the same category of the query image.

Some of the retrieved images for sample test images are given in Table I. The following performance measures were used to evaluate the performance of the algorithm.
Table 1: Outputs for Sample Test Images

<table>
<thead>
<tr>
<th>Test Image</th>
<th>Some Retrieved Images</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Images" /></td>
</tr>
<tr>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Images" /></td>
</tr>
<tr>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Images" /></td>
</tr>
<tr>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Images" /></td>
</tr>
</tbody>
</table>

Fig. 2(a): Proposed Algorithm

Fig. 2(b): K-Means Algorithm

Fig. 2, Plots of average values of $\alpha$ and $\beta$ against no. of clusters

Fig. 3(a): Proposed Algorithm vs K-Means Algorithm

Fig. 3(b): Proposed Algorithm vs K-Means Algorithm

Precision = \frac{\text{Total number of retrieved relevant images}}{\text{Total number of retrieved images}}

Recall = \frac{\text{Total number of retrieved relevant images}}{\text{Total number of relevant images}}
VII. Conclusion and Future Work
We have thus proposed a new partitional clustering algorithm based on the notion of ‘contribution of a data point’. Unlike the k-means algorithm, our algorithm optimized on both the intra-cluster and inter-cluster similarity measures and required fewer passes with each pass having the same time complexity as that of the k-means algorithm. We applied the clustering algorithm to content-based image retrieval and our experiments reveal that the algorithm improves on recall at the cost of precision. As with many other clustering algorithms, a limitation with our algorithm is that it requires the number of clusters to be known in prior. Various methods exist to determine the number of clusters for a given dataset [11] including the one based on game theory [10]. A problem with the k-means and k-medoids clustering algorithm is that they do not perform well when the clusters are non-spherical in shape. Whether the proposed algorithm suffers from such a limitation is yet to be investigated. Future lines of work would be to apply the concept of contribution to other clustering techniques such as hierarchical clustering. Our algorithm could also be extended to fuzzy partitioning of data points. The content-based image retrieval technique described in this paper uses only the RGB color histogram as the visual content descriptor of an image. The performance of the system with other visual features based on shape and texture and other distance metrics would have to be tested [8]. Also, learning through relevance feedback from the user could be incorporated in the proposed system [12].

References

![Fig. 3(c):](image-url) Fig. 3: Plots of average precision, Recall and F-measure against number of clusters. (fig a,b,c).

Note that an image is regarded as relevant if it shares at least one object common with the query image.

Since algorithms that improve on precision may degrade recall and vice versa, we use another performance evaluation measure called F-measure, which combines both precision and recall. F-measure is the harmonic mean of precision and recall and is given by

\[
F - \text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]


