

Evaluating Clustering Algorithms: Cluster Quality and Feature Selection in Content-Based Image Clustering

Mesfin Sileshi

Department of Computer Science
Addis Ababa University, Ethiopia
mesfin_sileshi2000@yahoo.com

Björn Gambäck

Department of Computer and Information Science
Norwegian University of Science and Technology
And: Swedish Institute of Computer Science
gamback@{idi.ntnu.no, sics.se}

Abstract

The paper presents an evaluation of four clustering algorithms: k -means, average linkage, complete linkage, and Ward's method, with the latter three being different hierarchical methods. The quality of the clusters created by the algorithms was measured in terms of cluster cohesiveness and semantic cohesiveness, and both quantitative and predicate-based similarity criteria were considered.

Two similarity matrices were calculated as weighted sums of a set of selected MPEG-7 color feature descriptors (representing color, texture and shape), to measure the effectiveness of clustering subsets of COREL color photo images. The best quality clusters were formed by the average-linkage hierarchical method. Even though weighted texture and shape similarity measures were used in addition to total color, average-linkage outperformed k -means in the formation of both semantic and cohesive clusters. Notably, though, the addition of texture and shape features degraded cluster quality for all three hierarchical methods.

1 Introduction

The straight-forward way to perform Content-Based Image Retrieval (CBIR) is to compare a query image to all images in a database. However, this is obviously computationally intensive and requires exhaustive search. Being a little smarter, we can either sort the images in the database into some pre-defined categories or cluster them together based on some criteria for similarity based on low-level features. This achieves scalability by avoiding exhaustive search; the query image initially needs to be compared to only a representative image from each category or cluster.

This study aims at measuring the effectiveness of clustering algorithms by measuring the quality of the clusters and to identify appropriate low-level image features for se-

lecting better similarity measures. The formation of quality clusters requires considering color, texture and shape features by assigning appropriate weight according to their discriminative power. The paper is laid out as follows: the section following describes the clustering algorithms and the utilized similarity measures, while Section 3 discusses the overall topic of Content-Based Image Clustering and some previous work in the field. Section 4 details the actual clustering experiments, including the data set and MPEG-7 features used in the present study. Finally, Section 5 sums up the discussion and draws conclusions from the results.

2 Clustering

In contrast to categorisation, clustering is traditionally considered as unsupervised learning, since members are assigned to a class without prior membership information. Clustering algorithms have been applied to a range of problems in a wide variety of research fields with the aim to identify interesting and hidden distributional patterns.

2.1 Clustering Algorithms

Hierarchical clustering builds a cluster hierarchy, in a top-down (divisive) or bottom-up (agglomerative) fashion: either starting with one cluster and recursively splitting the most appropriate clusters with regard to some similarity metric, or starting with one-point clusters and recursively merging similar clusters. Splitting/merging continues until a given stopping criterion (e.g., number of clusters) is met.

The fundamental criterion in Hierarchical Agglomerative Clustering Methods (HACM) is the identification of clusters that should be linked or combined. The most commonly used linkage methods are complete-link, average-link, and Ward's method [5, 16]. The *complete linkage* method measures the maximum inter-group distance among given pairs of groups and links those pairs that have the

smallest maximum separation. The method forms smaller, tighter and more compact clusters. The *average linkage* method is a mean method that links groups with the smallest average inter-group distance. *Ward's method* puts the emphasis on minimization of the information loss in each grouping in terms of the error sum-of-squares. The centroid of the temporarily merged groups is determined so that the average squared distance to the centeroid or the variance is computed. The merged groups with the smallest variance are linked, resulting in more homogeneous clusters.

k-means implements least-square partitioning in classifying an input data set into an *a priori* fixed number, k , of groups. The algorithm starts by initializing k cluster centers. The center is updated every time a new member is assigned to the cluster. The process ends when the cluster centers do not change. The method of choosing the initial centroids determines the final cluster result.

2.2 Cluster Quality Measures

The effectiveness of a clustering algorithm can be evaluated in terms of cluster and semantic cohesiveness [13, 8]. A highly *cohesive cluster* is one in which its members are more semantically similar to each other than to members of a different cluster. The best cohesive cluster is formed when all elements of a cluster belong to the same category.

Assume that a cluster c is formed by members from k different semantic categories, that n_i is the number of members from category i in c , and N the overall total number of elements in the cluster c (so $N = \sum_{i=1}^k n_i$). The probability of selecting a member of semantic category i in a cluster c is then given by $p_{ic} = n_i/N$ (and thus $\sum_{i=1}^k p_{ic} = 1$). The cluster cohesiveness (C) can be defined as follows:

Definition 1 (Cluster Cohesiveness)

$$C = - \sum_{i=1}^k p_{ic} * \log_2(p_{ic}) \quad (1)$$

When most members of a cluster belong to a specific category the measured value of the cluster cohesiveness becomes smaller. $C = 0$ for the best cohesive cluster, with all cluster elements coming from one category, while $C = 1.0$ implies a cluster consisting of members only from two categories with equal number from each. The maximum value is obtained for the worst case, with equal number of members from all categories assigned to a cluster.

A *semantic category* consists of a number of elements that are more semantically similar to each other. When a clustering method is applied to a specific semantic category, each of the clusters formed consists of members of the category in different proportion. If most members of a category belong to one specific cluster, the cohesiveness of the semantic category is said to be higher.

If a semantic category s consisting of M members is distributed over m clusters when a specific clustering method is applied, and n_{js} is the number of members from the category belonging to cluster j , then the probability of selecting a member of semantic category s in a cluster j is $p_{js} = n_{js}/M$. The information content of a semantic category in all the clusters reflects the cohesiveness of the category, and also indicates the separation of the semantics between the clusters. The cohesiveness of a category whose members are distributed over m clusters is measured as:

Definition 2 (Semantic Cohesiveness)

$$S = - \sum_{j=1}^m p_{js} * \log_2(p_{js}) \quad (2)$$

The measured value of the semantic cohesiveness becomes smaller, when most members of a category belong to a single cluster. When all members of a given category belong to only one cluster, the probability $p_{js} = 1$, and thus $S = 0$. The semantic cohesiveness of a category is worst when all its members are distributed equally in all clusters.

3 Content-Based Image Clustering

The first and basic step in content-based image clustering is the selection of the low-level features that are represented by unlabeled feature vectors. The image clustering process requires the selection among many different kinds of features, so that the images within a cluster are more similar to each other than images belonging to a different cluster.

3.1 Low-Level Features

The main goal of the feature extraction process is to reduce the low-level information of an image into a manageable amount of relevant properties. This reduces the complexity of the feature description process and makes the descriptors robust. The low-level features are grouped into local and global features. Global features represent the overall visual appearance of an image and describe it by means of histograms (color and texture) and overall layout. Local features extract local visual information (color, texture, and shape) partitioning the image into regions or objects.

The *color* property of an image is represented by a smaller set of color vectors and only the dominant features of the image color distribution, identifying regions that contain a predefined set of colors, normally represented by a color histogram. MPEG-7 specifies the number of bins in a color histogram (i.e., its discrimination power), the quantization scheme (i.e., uniform or non-uniform) together with the color space used [14]. MPEG-7 includes four color descriptors: the Color Layout Descriptor, the Dominant Color

Descriptor, the Color Structure Descriptor, and the Scalable Color Descriptor, with the last two being histogram-based.

Texture refers to the repetition of a basic pattern over a given area. The texture feature is defined in terms of the shape of the basic pattern and its repetition rate. Surfaces are described by textural properties that express structural arrangement and relative relationship to the surrounding environment. A texture feature has spatial extent, unlike a color feature which has point-wise (pixel) property. Images that have similar color content, for example sky and sea, are distinguished by their textural similarity.

The *shape* of a visual object is determined by the geometric relationship between points in an image. The basic assumption in shape analysis is considering shape of an object as characteristics of a binary image region. This allows the description of a shape to be computed from the object's boundary or its interior content. Shape descriptors can be divided into three main categories: area (region) based, contour (boundary) based, and skeleton-based descriptors [9].

3.2 Previous Work

Overall, CBIR performance needs to be enhanced. Previous studies have showed that the discriminative powers of feature descriptors varies, but that using MPEG-7 description allows interoperable search, retrieval and clustering.

CLUE, 'CLUster-based rEtrieval of images by unsupervised learning' [3] generated a cluster of images tailored to characteristics of a query image, and returned a set of images ranked according to similarity measures. A graph representation of the image collection was used to emphasize pair-wise relationships when measuring overall similarity, with the nodes of the graph representing the images. The edges were labeled with weights indicating the similarity. The limitations of CLUE included lack of cluster quality and poor semantic similarity among members; also, not all images in a database were considered.

Park *et al.* [12] suggested a method utilizing HACM-based clustering of retrieved images for re-ranking of the retrieval results. The similarity of the shape and color features was calculated from the sum of the absolute difference measure from the corresponding histograms, with Min/Max normalization applied to adjust the different range of similarity values from the color, texture and shape features before calculating the total similarity through a weighted sum of each feature. Only City-block distance measures were used and MPEG-7 descriptors were not considered.

Earlier, Abdel-Mottaleb *et al.* [1] incorporated the local features of the images in the database by dividing them into a fixed number of rectangular regions. The local variation of color information was captured by histograms and the similarity measure between the corresponding regions was combined to calculate a single similarity measure be-

tween images. Experiments on 200 photo images showed a larger reduction in the number of comparisons with clustering without sacrificing the retrieval accuracy. Two clustering algorithms implementing histogram intersection similarity measures performed better than k-means clustering.

The quality of the MPEG-7 descriptors (except for the region-based shape descriptor) has been analyzed by applying different feature extraction algorithms to image collections of three media types: monochrome texture, color photos from the Corel data sets (see Section 4.1), and a set of artificial color images with few color gradations [7]. The only media that performed well with all color descriptors was Corel photos. Except for the Dominant Color Descriptor, the color descriptors did not perform well neither for monochrome content nor for artificial media objects.

4 Experiments

The experiments were conducted in four phases. First, the actual set of images to be used as data was selected and evaluated. Then the low-level feature descriptors of each image were extracted, and quantitative and predicate-based distance criteria were used to measure similarity. Descriptor level similarity measures were combined to produce a total similarity matrix describing the distance of each image to all others. Finally, the clustering result of each of the algorithms was analyzed to measure the cluster quality.

4.1 Data Set Selection

The availability of *a priori* knowledge about the image data sets to be clustered is the basic requirement in cluster validity analysis, since it enables the measurement of the semantic cohesiveness of each category after implementing a specific clustering method [4, 8]. We chose semantically similar images from the COREL image database [15]. The images came from eight categories that were also used in SIMPLIcity [15] and CLUE [3]: 1. Mountains and glaciers, 2. Buses, 3. Foods, 4. Dinosaurs, 5. Elephants, 6. Flowers, 7. Buildings, 8. Africa people and villages.

Even though the set of images in a category were considered as similar in CLUE and SIMPLIcity, two groups of students were made to select semantically similar images among the 100 images in each category. The ways humans interpret the same image vary as perception varies among people. The students were neither informed about the category given by the database, nor able to compare other images to a predefined image template. Images that were selected by both student groups were used to form the data sets for the experiment. The students used different criteria of their own in selecting similar images: when 100 images of buses were provided, a subject differentiated two or more similar buses by their color. Most of the selected buses

were of the same dominant color and all were one-floored, similar-size buses. The direction of motion of the buses and the way the surrounding environment looked like were thus taken as other parameters in categorizing buses.

The images in each semantic category are stored in JPEG format with a size of 384x256 or 256x384 pixels. Among the images that were presented to the students, they selected equal-sized images, taking size as a factor to distinguish between the images' semantic similarities. A subject categorized semantically similar images as different based on their size. To avoid this, only images of size 384x256 were used in the actual experiments. The minimum number of commonly selected images from Categories 1, 2, 5 and 8 was 40 by both student groups. The commonly selected image numbers for the other four categories ranged from 45 to 65. To normalize category size, we took the most highly ranked images from each category. Therefore the total numbers of color photos that were used as ground truth information were 320 (40 from each of the 8 categories).

Normalization of category size has both advantages and disadvantages: while equally-sized categories is clearly not the common case in real-world problems, it assures that there is at least a certain number of items belonging to each category even in experiments with small data sets. Normalization also has the property of minimizing the baseline. Since the largest category is equal in size to the smallest one, simply guessing that an item belongs to the most common category has minimal effect on performance.

Using the entire COREL data set would have been the other option; however, measurement of the effectiveness of clustering methods does not necessarily require larger numbers of ground truth image data sets as efficiency is not a major issue, albeit it could be of importance for the cluster formation, for example, in order to avoid over-fitting of the cluster parameters. Most previous research has tried to cluster images aiming at improving retrieval performance and accuracy of results for an image query by users. When the image retrieval performance is the major issue it is necessary to consider large number of image data sets. Better accuracy is obtained as a result of the quality of the clusters formed by implementing the appropriate clustering method.

4.2 Features Considered

Before the beginning of the feature extraction process, each image was indexed randomly independent of its semantic category information. Each of the MPEG-7 image descriptors discussed in Section 3.1 extracts some properties from the visual media considered. Seven of them were used in representing the selected color image categories. All color descriptors: Color Layout, Color Structure, Dominant Color, and Scalable Color; two texture descriptors: Edge Histogram and Texture Browsing; and one shape descriptor:

Feature	Descriptor	Bins
Color	Color Structure (CSD)	32
	Scalable Color (SCD)	64
	Dominant Color (DCD)	57
	Color Layout (CLD)	12
Texture	Edge Histogram (EHD)	80
	Texture Browsing (TBD)	5
Shape	Region Shape (RSD)	35
	Contour-Based (CBD)	3

The color descriptors use different color spaces, as follows. CSD: HMMD, SCD: HSV, DCD: RGB, and CLD: YCrCb.

Region-based Shape. Descriptor extraction was performed using the MPEG-7-reference implementation using ACE-TOOLBOX [10] of M-OntoMat-Annotizer [2] that provides the extracted descriptors in XML format.

The bin values corresponding to the four MPEG-7 color descriptors were transformed into a matrix of 320 rows (the number of images in the experiment) and 165 columns (bin values of the descriptor elements). The two texture descriptors, Edge Histogram and Texture browsing were transformed into a data matrix of 85 columns for the same number of images. Finally, the feature vectors of the two shape descriptors were transformed into 38 columns. Therefore, each of the images was represented by a total feature vector of length 288. The MPEG-7 image descriptors extracted and the corresponding number of bins are shown in Table 1.

The number of bits used in representing a bin, that is, an element of a feature vector for each of the eight MPEG-7 descriptors is not the same for all bins. For example, the bin value of the Color Layout Descriptor ranges from 0 to 255 as 8 bits are used in representing each bin of the descriptor. The use of 3 bits in representing a bin of a Region Shape Descriptor allows a maximum value of 7 for its bin. Hence, before measuring the distance between image descriptors, it is necessary to column-wise normalize the resulting values for each vector element, so that it will be in the range [0, 1].

In order to evaluate the effectiveness of the image clustering, cluster and semantic cohesiveness were measured (Section 2.2). The quantization model [6] was used in computing the distance between each of the image descriptor with all other images. The similarity measures used in measuring similarity among the extracted MPEG-7 feature descriptors for the four color descriptors are from both the quantitative and predicate-based similarity measure domain. Pattern difference (a predicate-based distance measure) was implemented for CSD, SCD, EHD, and RSD, while Meehle Index and Clark's divergent coefficient were used as quantitative similarity measures for DCD and CLD, and the Pearson correlation coefficient for measuring distance for TBD.

Table 2. Cluster cohesiveness: color

Cluster	k-means	Avg.	Compl.	Ward
1	1.4355	2.7652	0.0000	0.7532
2	0.7025	0.0000	2.3156	1.7695
3	1.4859	0.4310	0.4537	2.3492
4	2.1713	0.0000	2.2384	0.6582
5	1.7324	0.0000	1.5739	0.4971
6	2.6547	0.2007	0.0000	0.1720
7	1.3792	1.0000	1.2171	0.0000
8	1.5610	0.0000	0.6581	0.0000

Table 3. Cluster cohesiveness: total

Cluster	k-means	Avg.	Compl.	Ward
1	2.6630	0.6889	2.7773	1.2580
2	0.7344	2.7119	0.0000	2.0875
3	1.5510	1.0000	2.1747	2.2738
4	0.4453	1.0000	1.0000	0.6052
5	1.5218	0.0000	1.4825	1.0782
6	2.1426	0.0000	1.0000	1.0000
7	1.1730	1.0000	1.4825	1.0000
8	1.2980	1.0000	2.3220	0.0000

The combined texture descriptor similarity needs to assign more weight to EHD due to its higher discriminative power [6]. Shape was represented using only RSD since the other basic shape descriptor, CBD gave the same values for 98% of the images under consideration.

A previous empirical evaluation by Ojala *et al.* [11], of the four MPEG-7 color descriptors showed DCD to be worst in retrieving semantic image categories. Using a combined color similarity measure of all descriptors by assigning more weight to CSD provided the retrieval with the most accurate semantic categories. Wong *et al.* [17] tried to boost the performance of retrieval based on DCD by combining it with CSD, in order to achieve the compactness of the former descriptor and the accuracy of the latter.

In the present study, overall image similarity is calculated by summation of the combined color, texture, and region-based shape similarity measures. The relative importance of the features are not the same, so they are weighted:

$$D_{combined} = w_c * D_{color} + w_t * D_{texture} + w_s * D_{shape} \quad (3)$$

In deciding on the weight for each of the three features, the optimal results, obtained after several tests, were based on the bin ratio (the total number of bins of a feature divided by the total number of bins for all three features). For the experiments in this paper the weights in (3) thus were:

$$w_c = 0.45, w_t = 0.35, w_s = 0.20 \quad (4)$$

The sections following discuss the cluster cohesiveness and semantic cohesiveness measured as a result of the k-means and the three agglomerative hierarchical clustering methods, in terms of the results obtained when the *total color similarity* and *total feature similarity* matrices were input to the four clustering algorithms.

For a category in a worst case cluster, Equation 1 gives a value of $C = 0.375$, so the total sum for a cluster is $0.375 * 8 = 3.0$. Also, when all members of a category are distributed equally in all clusters, Equation 2 gives 3.0 as maximum. Thus both cluster cohesiveness and semantic cohesiveness are in the range $[0.0, 3.0]$ in this experiment.

4.3 Evaluating Cluster Cohesiveness

Table 2 shows the cluster cohesiveness results obtained for each algorithm when the total color similarity matrix is input, while Table 3 shows the results when the input is the total similarity matrix (which incorporates the total color similarity). Hence the value 0 in the color-based clusters corresponds to a value ranging from 1 to 32 images of the same category in a cluster. This value ranges from 4 to 32 for the total similarity cluster cohesiveness.

Four of the clusters formed by the average HACM in the color-based clustering contain images from one category each; three clusters contain a maximum of three different semantic categories, while the last cluster is formed from all the categories. This makes the average HACM the best method in forming cohesive clusters. The results for complete HACM and Ward are similar to each other; however, the cluster cohesiveness by the complete method with values > 1.0 show that four of the clusters are formed from images of four or more categories, while Ward produces only two such clusters, and thus performs better. Most of the clusters formed by k-means are composed of images from more than four categories; only one of its measured values is < 1.0 , making k-means the worst method in the formation of cohesive total color-based clusters.

When feeding the total similarity matrix to the algorithms (Table 3), the worst cluster formed by k-means consists of 47 images from all eight categories. The best cohesive clusters with measured values < 1.0 consist of images from three categories. Average HACM gives a more uniform distribution of semantic categories: Two of the clusters formed by it consist of images from one category, but it also forms the worst compact cluster with 226 images from seven categories. The complete HACM is the worst in forming cohesive clusters: five of the clusters are made up of images from 5-8 categories. Cluster cohesiveness values > 2.0 imply clusters containing seven categories or more. Two of the other clusters contain equal number of images from two categories. Ward's method is relatively better than complete HACM in total feature-based image clustering.

Table 4. Semantic cohesiveness: color

Category	k-means	Avg.	Compl.	Ward
1. Mountains	0.7625	0.4531	1.4890	0.8485
2. Buses	1.3762	0.8305	0.6954	0.8539
3. Foods	1.9287	1.0368	1.6148	1.2192
4. Dinosaurs	0.2864	0.0000	0.2864	0.2864
5. Elephants	1.9584	0.2689	1.8005	1.0638
6. Flowers	1.3414	0.8822	0.1686	0.9284
7. Buildings	2.0958	0.1686	0.8538	0.6097
8. Africa...	1.6128	0.1686	1.0120	1.5229

Table 5. Semantic cohesiveness: total

Category	k-means	Avg.	Compl.	Ward
1. Mountains	0.4531	0.8286	2.2452	1.1320
2. Buses	1.1498	0.4531	2.0144	1.8935
3. Foods	1.1935	0.8286	2.2452	1.1320
4. Dinosaurs	0.9414	0.4531	2.0144	1.8935
5. Elephants	2.2429	0.2689	1.4001	1.3382
6. Flowers	1.0295	0.7218	0.8112	0.7691
7. Buildings	1.9724	0.0000	0.6908	0.5474
8. Africa...	1.5476	0.0000	1.0869	1.1884

4.4 Evaluating Semantic Cohesiveness

Tables 4 and 5 show the measured semantic cohesiveness of each category, when the total color similarity and total similarity matrices are input, respectively.

Generally, all the three HACM methods result in better semantically cohesive clusters than k-means for the same total color similarity input matrix (Table 4). The average method is best among all as its worst result partitions each of four of the high-valued categories into four groups each. The complete and Ward methods result in nearly similar semantic cohesiveness values for most categories, but the latter is somewhat better. Interestingly, the semantic cohesiveness of Category 4 (Dinosaurs) is best for all clustering methods except for the complete HACM, for which it is second best. The average and complete HACMs are not good in creating cohesive semantics of Category 3 (Foods). Ward produces relatively better cohesive Category 3 semantics, but is worse with respect to other categories. k-means partitioned each of the six image categories with cohesiveness > 1.0 into 4-6 parts. It is worst for Categories 5 and 7 where the members are partitioned into six different clusters.

When the total similarity matrix is used as input (Table 5), the three HACM produce the best cohesive semantics of Category 7. Even though the complete and Ward methods distribute member images of Category 7 in three different clusters, about 80% of the images of each category belonged to one specific cluster. Similarly, all of the HACM methods give the next best semantic cohesive values for Category 6, partitioning it into two parts only. The semantic cohesiveness of Categories 1 and 3 are equally worst for average and complete HACM: the complete method distributes the members of each of these categories on average in six different clusters. The complete method is not suitable for forming cohesive clusters in the case of the first four categories, each with measured values > 2.0 . This implies a distribution of each category into six different clusters, ranking complete HACM as worst, while the measured values of the average method all are < 1.0 . Ward’s method distributes four categories into three clusters, while k-means distributes four categories into more than four clusters.

4.5 Overall Cohesiveness Measures

The overall measures of cluster and semantic cohesiveness of each of the total color-based and total feature-based clustering with respect to the four clustering methods are shown in Table 6. When m clusters are formed applying a specific clustering method, the total cluster cohesiveness is measured as the sum of the cohesiveness of each cluster: $-\sum_{c=1}^m \sum_{i=1}^k p_{ic} * \log_2(p_{ic})$. And when k semantic categories are used for the formation of m clusters, the total semantic cohesiveness is measured as the sum of the cohesiveness of each category: $-\sum_{s=1}^k \sum_{j=1}^m p_{js} * \log_2(p_{js})$. Thus, the values for the Cluster and Semantic cohesiveness measures for each method in Table 6 range from 0.0 to 24.0 (and accordingly, their Sum from 0.0 to 48.0).

All three HACM methods show decrease in total cluster cohesiveness with the addition of texture and shape features information to the total color similarity. While in contrast, k-means shows an improvement in the overall cluster cohesiveness in total feature-based clustering.

For semantic cohesiveness, adding texture and shape information to the total color similarity matrix was expected to improve the overall cohesiveness of the categories, but the results show that there is no significant improvement. The best improvement in the measured values is < 1.0 , while the texture and shape similarity information actually significantly *degrade* the semantic cohesiveness of both the complete linkage HACM method and of Ward’s method.

Table 6. Overall cohesiveness measures

	k-means	Avg.	Compl.	Ward
Cluster				
Color	13.1225	4.3969	8.4568	6.1992
Total	11.5297	7.4008	13.4302	9.3027
Semantic				
Color	11.3622	3.8087	7.9205	7.3328
Total	10.5302	3.5541	12.5082	9.8941
Sum				
Color	24.4847	8.2056	16.3773	13.5320
Total	22.0599	10.9549	25.9384	19.1968

5 Conclusions

The paper addressed the problems of clustering color photo images based on their low-level feature representation. We considered k-means clustering and agglomerative hierarchical clustering using three different cluster linkage methods. Even if clustering results are affected by the selection of features and proximity measures, the study carefully used the best in the selection process in a way that the algorithms were the only discriminating factor.

Among the three HACM used for the total color similarity matrix input, the best quality clusters were formed by the average-linkage method. The quality obtained by it was twice that of complete HACM. Ward's method produced clusters of similar quality, but better than the complete method. The quality of clusters formed by k-means was worse than all three hierarchical methods in using total color similarity, and three times less than the best HACM.

The addition of texture and shape similarity measures (relatively in a lesser weight to the total color similarity) provided a different result. Similar to the total color-based clustering, the average HACM was the best method compared to both k-means and the other two hierarchical methods, and both in terms of semantic and cluster cohesiveness. Even though the quality of clusters using Ward's method was half of the average method, it resulted in more cohesive clusters than k-means clustering, which in turn outperformed the complete hierarchical method.

Generally, hierarchical agglomerative clustering with average-linking gave more than twice as good quality clusters as those of the k-means method regardless of whether total color or total feature similarity was used. Only using the complete HACM resulted in k-means being better than a hierarchical method in the formation of quality clusters.

The other interesting point is that instead of the expected improvement in cluster quality of color photos, the addition of texture and shape feature degraded cluster quality for all hierarchical methods, while using the overall total image similarity matrix resulted in a significant improvement for the k-means method.

6 Acknowledgments

This work was partially supported by Sida, Swedish International Development Cooperation Agency through SPIDER (Swedish Programme for ICT in Developing Regions), and by the European Commission's Information Society Technologies programme, contract IST-FP6-34434, 'COMPANIONS' (www.companions-project.org).

Thanks to Dr. Fredrik Olsson (SICS), Kibur Lisanu, Samuel Eyassu, Lemma Nigusie, and Firehiwot Sahlu for valuable help and support, as well as to Dr. Solomon Atnafu (Addis Ababa University) for co-supervision.

References

- [1] M. Abdel-Mottaleb, S. Krishnamachari, and N. J. Mankovich. Performance evaluation of clustering algorithms for scalable image retrieval. *Empirical Evaluation Techniques in Computer Vision*, pp. 45–56, Santa Barbara, CA, 1998. Wiley-IEEE Computer Society.
- [2] S. Bloehdorn, K. Petridis, C. Saathoff, N. Simou, V. Tzouvaras, Y. Avrithis, S. Handschuh, Y. Kompatsiaris, S. Staab, and M. G. Strintzis. Semantic annotation of images and videos for multimedia analysis. *2nd Europ. Semantic Web Conf.*, pp. 592–607, Heraklion, Crete, 2005.
- [3] Y. Chen, J. Z. Wang, and R. Krovetz. CLUE: Cluster-based retrieval of images by unsupervised learning. *IEEE T. Image Process.*, 14(8):1187–1201, 2005.
- [4] P. Cunningham and R. Dahyot. A review of machine learning techniques for processing multimedia content. MUSCLE Deliverable D8.1, Dublin, Ireland, 2004.
- [5] C. Ding and X. He. Cluster merging and splitting in hierarchical clustering algorithms. *2nd Int. Conf. Data Mining*, pp. 139–146, Maebashi, Japan, 2002. IEEE.
- [6] H. Eidenberger. Distance measures for MPEG-7-based retrieval. *5th Int. Workshop Multimedia Information Retrieval*, pp. 130–137, Berkeley, CA, 2003. ACM SIGMM.
- [7] H. Eidenberger. Statistical analysis of MPEG-7 image descriptions. *Multimedia Syst.*, 10(2):84–97, 2004.
- [8] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3):107–145, 2001.
- [9] J. Latecki, R. Lakämper, and D. Wolter. Shape similarity and visual parts. *11th Int. Conf. Discrete Geometry for Computer Imagery*, pp. 34–51, Naples, Italy, 2003.
- [10] N. E. O'Connor, E. Cooke, H. L. Borgne, M. Blighe, and T. Adamek. The AceToolbox: Low-level audiovisual feature extraction for retrieval and classification. *2nd Europ. Workshop Integration of Knowledge, Semantic and Digital Media Technologies*, pp. 55–60, London, England, 2005.
- [11] T. Ojala, M. Aittola, and E. Matinmikko. Empirical evaluation of MPEG-7 XM color descriptors in content-based retrieval of semantic image categories. *16th Int. Conf. Pattern Recognition*, vol. 2, pp. 1021–1024, Quebec, 2002. IEEE.
- [12] G. Park, Y. Baek, and H.-K. Lee. Re-ranking algorithm using post-retrieval clustering for content-based image retrieval. *Inform. Process. Manag.*, 41(2):177–194, 2005.
- [13] M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. *IEEE T. Pattern Anal.*, 29(1):167–172, 2007.
- [14] T. Sikora. The MPEG-7 visual standard for content description. *IEEE T. Circ. Syst. Vid.*, 11(6):696–702, 2001.
- [15] J. Z. Wang, J. Li, and G. Wiederhold. SIMPLiCity: Semantics-sensitive integrated matching for picture libraries. *IEEE T. Pattern Anal.*, 23(9):947–963, 2001.
- [16] J. H. Ward, Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.*, 58(301):236–244, 1963.
- [17] K.-M. Wong, L.-M. Po, and K.-W. Cheung. Dominant color structure descriptor for image retrieval. *Int. Conf. Image Process.*, vol. 6, pp. 365–368, San Antonio, TX, 2007. IEEE.