A CLUSTERING-BASED APPROACH FOR EVALUATION OF EO IMAGE INDEXING

Reza Bahmanyar[a]*, Gerhard Rigoll[b], Mihai Datcu[c]

[a] Munich Aerospace Faculty, German Aerospace Center (DLR),
Oberpfaffenhofen, 82234 Wessling, Germany - gholamreza.bahmanyar@dlr.de

[b] Munich Aerospace Faculty, Institute for Human-Machine Communication, Technical University of Munich,
Theresienstr. 90, 80333 Munich, Germany - rigoll@tum.de

[c] Munich Aerospace Faculty, German Aerospace Center (DLR),
Oberpfaffenhofen, 82234 Wessling, Germany - mihai.datcu@dlr.de

KEY WORDS: Clustering, Internal cluster indexing, External cluster indexing, Information Retrieval systems, Feature extraction, Earth Observation

ABSTRACT:

The volume of Earth Observation data is increasing immensely in order of several Terabytes a day. Therefore, to explore and investigate the content of this huge amount of data, developing more sophisticated Content-Based Information Retrieval (CBIR) systems are highly demanded. These systems should be able to not only discover unknown structures behind the data, but also provide relevant results to the users' queries. Since in any retrieval system the images are processed based on a discrete set of their features (i.e., feature descriptors), study and assessment of the structure of feature space, build by different feature descriptors, is of high importance. In this paper, we introduce a clustering-based approach to study the content of image collections. In our approach, we claim that using both internal and external evaluation of clusters for different feature descriptors, helps to understand the structure of feature space. Moreover, the semantic understanding of users about the images also can be assessed. To validate the performance of our approach, we used an annotated Synthetic Aperture Radar (SAR) image collection. Quantitative results besides the visualization of feature space demonstrate the applicability of our approach.

1 INTRODUCTION

Intensively increasing the volume of Earth Observation (EO) data in recent years, in order of several Terabytes a day, increases the demand to develop more sophisticated Content-Based Information Retrieval (CBIR) systems. These systems facilitate exploring the content of available datasets to provide relevant results to the users' queries. In spite of large amount of research in this area, the results provided by CBIR systems are not always satisfactory. Because, the users' queries are usually based on high-level semantics (e.g., river, building, forest), but retrieval systems explore image collections based on their descriptions by a discrete set of low-level features (e.g., shape, texture, color). Therefore, investigation and study of feature descriptors and the topology of the resulted feature space is a big challenge in information retrieval process chain. This study helps to develop more sophisticated feature descriptors which can categorize the given collection of images to semantically meaningful categories or be more specific to classify a particular semantic category based on the users' queries.

In our paper, we deal with two main challenges of CBIRs (e.g., image feature extraction and data annotation) to not only facilitate the annotation procedure for the users, but also make CBIRs able to provide user satisfactory results. To this end, we investigate the structure of the feature space, built by different feature descriptors (e.g., Gabor, WLD, Rand_Feat), using a clustering-based evaluation technique. The main idea is to cluster the given image collection for different number of clusters, then evaluate the clusterings both internally and externally. While internal evaluation allows to find the optimum clusters by relying on the unsupervised nature of clustering (i.e., without using prior annotations), external evaluation is a way to compare the resulted clusters to the users understandings of images (i.e., prior annotations). Finding the optimum clusterings for different kinds of feature descriptors indicates the homogeneity and discriminability of the images in the sense of different features (e.g., shape, texture, color), then comparing to the provided annotation indicates how these features are considered by the users to group or discriminate the images.

While previous works relay on either internal evaluation of clusterings, e.g., (FÃrber et al., 2010, Halkidi and Vazirgiannis, 2001), or the external evaluation using prior annotations, e.g., (Vinh et al., 2010, Larsen and Aone, 1999), we claim both evaluations are essential to investigate and explore the structure of feature space. The unsupervised nature of clustering allows exploring feature space regardless of annotation. Therefore, clustering-based assessment can be used before and after annotation to ease and validate the annotation procedure. While internal evaluation provides the users by an overview of the structure of the given image collection, the external evaluation validates the annotated classes.
For demonstration of our approach, an annotated collection of SAR images is used. Then quantitative results as well as visualizations of feature space for three different feature descriptors are provided to show how internal and external evaluations can provide knowledge about the structure of feature space.

The rest of this paper is organized as follows. Section 2 provides an overview of cluster evaluation as well as the used internal and external evaluation techniques. In Section 3, we give an introduction of our clustering-based evaluation approach. Section 4, provides an overview of the used dataset and the three feature descriptors. Moreover, the experimental results demonstrate the performance of our approach. Finally, in Section 5, we conclude our work.

*Corresponding author. gholamreza.bahmanyar@dlr.de, +49 815328 1803

2 INTERNAL AND EXTERNAL EVALUATION OF CLUSTERING

The unsupervised nature of clustering makes validation of resulted set of clusters a big challenge. Since the fundamental purpose of using unsupervised methods such as clustering is to distinguish unknown structure of data, validation without having external knowledge such as predefined class structures is highly demanded. Moreover, in real world data, due to the unknown structures behind the data, the provided annotation is not sufficient for validation of clustering (FÃrber et al., 2010). Therefore, the annotation does not always correspond to the natural grouping of the data points. There are variety of methods proposed to evaluate clustering internally (i.e., regardless of prior annotation) such as S_Dbw validity index (Halkidi and Vazirgiannis, 2001), Calinski-Harabasz (CH) (Caliński and Harabasz, 1974), Davies-Bouldin index (DB) (Davies and Bouldin, 1979), etc. Moreover, there are some previous works which compare these evaluation techniques from different aspects (Liu et al., 2010, Rendón et al., 2011). According to the reported comparison and some experiments done by ourselves, we found that S_Dbw can provide better evaluation of grouping procedure than the other techniques. This method consider both compactness and separability of discovered clusters at the same time which are the two main desired criteria in clustering methods (Tan et al., 2005). Moreover, based on the evaluation reported in (Liu et al., 2010), the indexing provided by S_Dbw is rather stable against monotonicity, noise, density, subclusters, and skewed distributions. Therefore, we use S_Dbw in our data assessments in this paper.

In addition to internal evaluation, we use external cluster indexing to explore the closeness of the available prior annotation to the clusters. This allows to compare the structure discovered by clustering to the understanding of users from data. Among several methods introduced in literature for external cluster indexing (e.g., Adjusted Random Indexing (ARI) (Hubert and Arabie, 1985), Adjusted Mutual Information (AMI) (Vinh et al., 2010), F-measure (Larsen and Aone, 1999)), we use ARI which is proved to be able to provide reasonable comparison of the prior annotation and the clusters (Vinh and Epps, 2009).

2.1 S_Dbw measure

Internal validation of clusters allows to choose the optimal clustering which fits the data best without any prior reference labeling. The main task of clustering techniques is to partition a given set of points in such a way that similar points goes to the same cluster, whereas the points in different groups are distinct.
S_Dbw is an internal validation technique which consider the both important criteria in clustering, e.g., within class similarity and between class distinguishablity (Halkidi and Vazirgiannis, 2001). This method computes the average scattering for clusters as a measure for intra-cluster similarity of points as the following,

$$Scat(C) = \frac{\frac{1}{c}\sum_{i=1}^{c} ||\sigma(v_i)||}{||\sigma(C)||}, \quad (1)$$

where c is the number of clusters, $||\sigma(v_i)||$ is the variance of cluster v_i, and $||\sigma(C)||$ is the variance of the whole data set.
In order to measure the distinguishability of clusters, S_Dbw computes the average density between clusters as the following,

$$Dens(C) = \frac{1}{c.(c-1)} \sum_{i=1}^{c} [\sum_{j=1, j\neq i}^{c} \frac{f(v_i, u_{ij}) + f(v_j, u_{ij})}{max\{f(v_i, vc_i), f(v_j, vc_j\}}], \quad (2)$$

where $f(v_i, .)$ is the number of points grouped in cluster v_i which lie on a certain distance of a point (e.g., point u_{ij} which is the middle point on the line connecting the two cluster centers vc_i and vc_j). In our experiments, the distance is taken equal to the average variance of the data set.
Combinig the average scattering and the density between clusters, S_Dbw is computed as the following,

$$S\_Dbw(C) = Scat(C) + Dens(C). \quad (3)$$

2.2 Adjusted Rand Index measure

Comparing the different grouping of a set of data points has been always a challenge in clustering methods. Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) is a partition comparing method base on a co-occurrence matrix of size $n \times m$, where m and n are the number of clusters in the two given clusterings $V = \{v_1, v_2, ..., v_n\}$ and $U = \{u_1, u_2, ..., u_n\}$, respectively. The values in this matrix show the number of co-occurrences of the data points in the clusters. Basically, this method measures how different pairs of data points are treated in the clustering procedure, i.e., how many pairs of points grouped to gather in the both clusterings (GG), how many of them are separated by both of the clusterings (SS), and how many of them are grouped in one clustering while separated in the other one (GS). Using these values, one computes the Rand Index (RI) as the following,

$$RI(U, V) = \frac{GG + SS}{GG + SS + GS}. \quad (4)$$

However, the expected value of the RI is changing in every experiments for even two partitions which leads to unfair comparison of the clusterings.Therefore, the adjusted version of RI , so called Adjusted Rand Index (ARI), is introduced in (Hubert and Arabie, 1985) as the following,

$$ARI(U, V) = \frac{RI(U, V) - Expected\ Index}{Max\ RI(U, V) - Expected\ Index}. \quad (5)$$

In ARI, the *Expected Index* is the expected value computed for (GG + SS) in a fixed experimental setup, e.g., for the two clusterings, in a random subject, the original number of clusters and data points in each is considered. Moreover, the maximum RI, $Max\ RI(U, V)$, is equal to 1.
ARI is equal to 1 as the two clusterings are identical and is equal to 0 when the RI of the two clusterings is equal to the *Expected Index*.

3 A CLUSTERING-BASED APPROACH FOR EVALUATION OF EO IMAGE INDEXING

During recent years understanding the structure of the information provided by available sources (e.g., Earth Observation, Multimedia, Biomedical) has become highly interesting. Because, data understanding is a fundamental step in developing retrieval, classification, and categorization systems. The data, in our case collections of EO images, is described by a discrete set of its representative features to be processed by computers. These feature descriptors, which can represent different aspects (e.g., shape, texture, color), provide computers with the low-level understanding of the structure of images. However, users' understandings of given images are usually based on high-level concepts. Consequently, usually the provided results by retrieval and categorization methods are not relevant to the users' queries (Bahmanyar and Datcu, 2013). Moreover, usually the provided annotations by the users cannot reflect the whole structure of data (FÃrber et al., 2010). In other words, the data provides knowledge from variety of aspects, while the users annotate images only from some

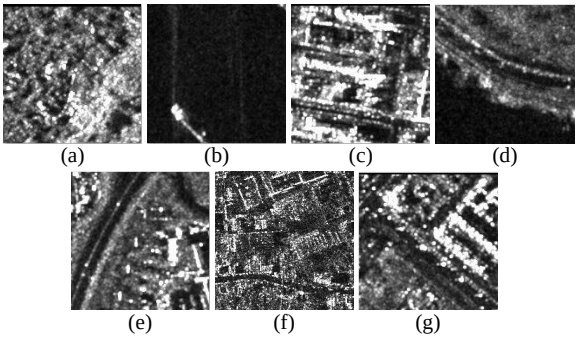


Figure 1: EO data set of 1230 SAR images grouped in seven non-equal size classes. (a) forest, (b) water, (c) medium density urban area, (d) forest + water, (e) roads, (f) high density urban area, and (g) urban area + roads).

aspects which let the other aspects to be still unknown. Therefore, studying the influences of different feature descriptors in understanding of the structure of data by computers is highly demanded.

In our paper, we explore the structure of the feature space, built by the feature descriptors extracted from EO images, by internal evaluation of clustering performed on the data. The unsupervised nature of clustering allows computers to find the structure of the feature space without being affected by any prior knowledge. In our experiments, we apply K-means, as the clustering technique, to a collection of SAR images for different number of clusters. Then we evaluate the clusters using an internal cluster evaluation method, namely S_Dbw (Halkidi and Vazirgiannis, 2001), to find the optimum clusters. The desired clusters are compact with a small density of feature points between clusters. Since there is no prior labeling considered in internal cluster evaluation, the clustering is not penalized for discovering new clusters or finding a different structure to the annotation labels (FÃrber et al., 2010).

Furthermore, we perform an external cluster evaluation, namely ARI (Hubert and Arabie, 1985), to compare the clusters to the prior annotation. This shows the difference between the understanding of the images by computers and the users' perceptions. Moreover, one can study from which aspects users annotating the given collection of images.

4 RESULTS AND DISCUSSION

4.1 EO image data set

In our experiments, we use a collection of EO data contains 1230 Synthetic Aperture Radar (SAR) images of size 160×160, Figure 1. The images are grouped in seven non-equal size classes (e.g., forest (198 images) , water (210 images), medium density urban area (204 images), forest + water (114 images), roads (67 images), high density urban area (279 images), and urban area + roads (158 images)). The images in the classes are rather homogeneous which allow to study the difference between the annotation and the resulting clusters.

4.2 Feature descriptors

In order to be processed by computers, the images are represented by a discrete set of extracted features. In this paper, we study the structure of feature space provided by three different feature descriptors, e.g., Gabor, WLD, and Random Features. In the feature space, each image is represented by a feature vector extracted from the entire image.

Gabor wavelet descriptors, proposed for texture analysis, are achieved by filtering a given image using a set of linear band-pass filters, the so called Gabor filters (Manjunath and Ma, 1996). These filters are generated by scaling and rotating a mother wavelet filter. The impulse response of this filter is a 2D modulated Gaussian function. The final Gabor feature vector is constructed by using the means (μ_{sr}) and the standard deviations (σ_{sr}) of the image filtered by S number of scales and R number of rotations, $F_{Gabor} = [\mu_{11}\ \sigma_{11}\ \mu_{12}\ \sigma_{12}\ ...\ \mu_{SR}\ \sigma_{SR}]$.
In our experiments, the Gabor features are constructed for 3 scales and 6 rotations which leads to a vector of 36 dimensions.

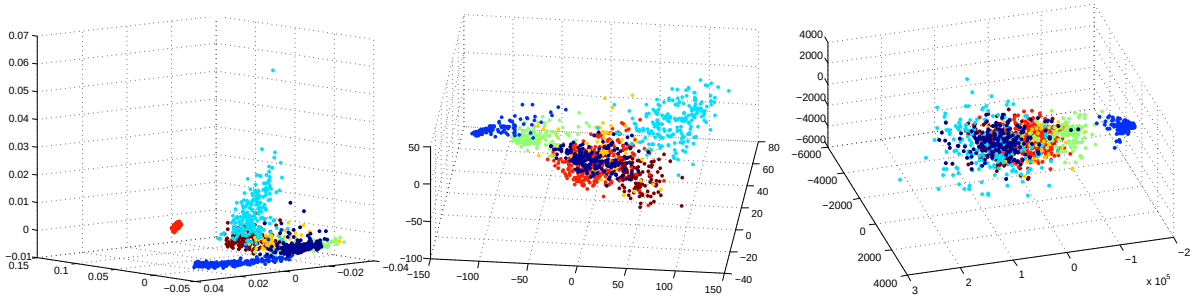
WLD is a feature descriptor developed based on Weber's law, a psychological law (Chen et al., 2010). According to this law, human notices the change in a stimuli as a valid signal if its ratio to the original intensity of the stimuli is above a certain constant value. WLD is constructed by a 2D histogram of: 1) Differential Excitation, the ratio between intensity difference between each pixel x and its neighbors; 2) Orientation, which is the gradient orientation of each pixel x. The final feature vector is constructed by building a 1D histogram of the computed 2D histogram, after quantizing to M number of excitations and T number of orientations.
In our experiments, we set M and T equal to 6 and 8, respectively, which results in a feature vector of 144 elements.

In order to use directly the intensity values, usually histogram of pixel values is constructed (e.g., color histogram in multimedia color images). Since the range of the intensity values of SAR data is rather wide which results in a very large vector, constructing the histogram of the pixel values is not trivial. Thus, in our experiments, we put all the pixel values of each given image $I_{m \times n}$ in a vector of size $d = m \times n$. Although in this way the size of the resulted vector is smaller, it is still too large to be used efficiently by the clustering techniques. Therefore, we used the idea of **Random Features (Rand_Feat)** (Liu and Fieguth, 2012) to decrease the dimensionality of the resulted feature vector d to a lower dimensional vector of size $\tilde{d}$. In this method, we compute the product of the high dimensional feature vector to a $d \times \tilde{d}$ random matrix.
In our experiments, we decrease the dimensionality of feature vectors to 32.
Figure 2 shows the 3D visualization of feature space built by the three feature descriptors. The prior annotation is also illustrated by different colors.

4.3 Internal evaluation of clusters

Internal evaluating of the clusters allows to investigate the structure of the given data, represented by feature descriptors, regardless of any prior knowledge.
In our experiments, S_Dbw is used to internally evaluate the clustering on three different feature descriptors (e.g., Gabor, WLD, and Rand_Feat) for different number of clusters (Figure 3c). Since S_Dbw is achieved by combining the average scattering and the average density between clusters, we show the two values as well in Figures 3a and 3b. As the results show, scattering monotonically decreases by increasing the number of clusters; however, after a certain number of clusters the change is not significant. Comparing the three feature descriptors, scattering decreases more for Rand_Feat than Gabor and WLD which means the structure of feature points is sparser in Rand_Feat.

As Figure 3b shows, average density between clusters does not change monotonically by increasing the number of clusters. Moreover, the general behaviors of the curves are rather different for different feature descriptors. It means that the average density



(a) WLD (b) Gabor (c) Random Feature

Figure 2: 3D visualization of feature space for three kinds of feature descriptors (e.g., WLD, Gabor, Rand_Feat) including the prior annotation of the data. Different colors represent different prior clusters (e.g.,forest, water, medium density urban area, forest + water, roads, high density urban area, and urban area + roads).

between clusters highly depends on the structure of the feature space. The optimum clustering provides clusters with minimum density of the points between clusters. For WLD we have the minimum density for two clusters which clearly indicates the feature space should be structured with two highly separated mass of points (Figure 4a). Furthermore, by increasing the number of clusters to three, the average density significantly increases due to splitting one of the separated masses into two clusters which come up with a large interfacing region (illustrated in Figure 4b); however, increasing the number of clusters to four, decreases the average density to some extent, as a newly added cluster introduces a small interface to the other clusters which leads to a decrease in the average density (illustrated in Figure 4c). For Gabor and Rand_Feat the behaviors of the curves provide general intuitions about the structure of corresponding feature spaces as well (some illustrations are shown in Figures 4d to 4i).

Combining the two criteria, average scattering and average density between clusters, results in S_Dbw measure. As it is illustrated in Figure 3c, for WLD and Gabor the optimality does not necessarily increases by increasing the number of the clusters; they have multiple local minimum which means there are more than one optimum clustering (e.g., for WLD we have optimum in 2, 4, 9, and 13 clusters). However, the behavior is rather different for Rand_Feat. The S_Dbw monotonically decreases which demonstrates that the feature space is not well structured leading to have the optimum number of clusters equal to the number of feature points.

4.4 External evaluation of clusters

As the results for internal clustering show, the optimum number of clusters is not necessarily equal to the number of the prior classes. Moreover, comparing the clusters, Figure 4, and the prior classes, Figure 2, illustrates that the points in one class does not group necessarily in one cluster. This demonstrates that the human semantic annotation does not necessarily correspond to what feature descriptors represent from the data. Therefore, we perform an external evaluation of clusters, which allows to compare the clusters to the prior annotations, using ARI. As Figure 5 illustrates, generally the structure represented by WLD and Gabor are closer to what users percept from the images in annotation time. In other words, users discriminate the images mostly based on the textures than the average intensity values. Moreover, comparing Figures 3c and 5, illustrates that the highest similarity of the clusters and the prior annotation does not necessarily occur at the optimum clusters.

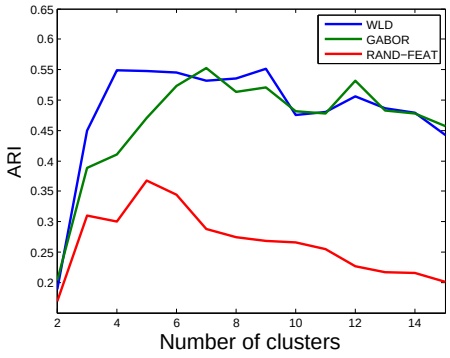
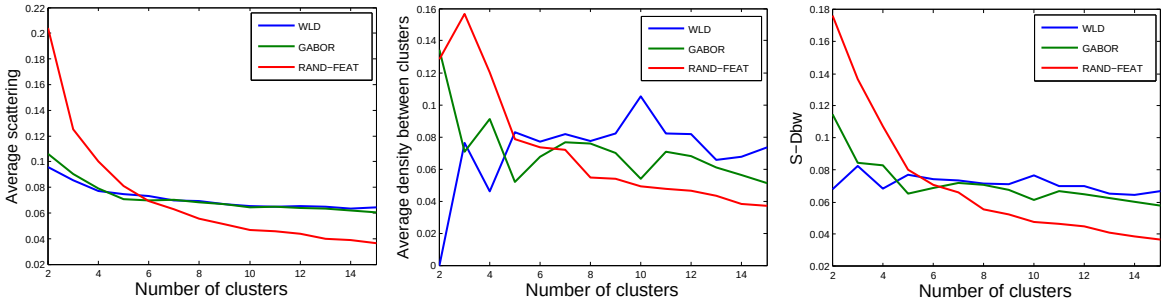


Figure 5: Adjusted Rand Index for three feature descriptors (e.g., WLD, Gabor, Rand_Feat).

5 CONCLUSIONS AND FUTURE WORK

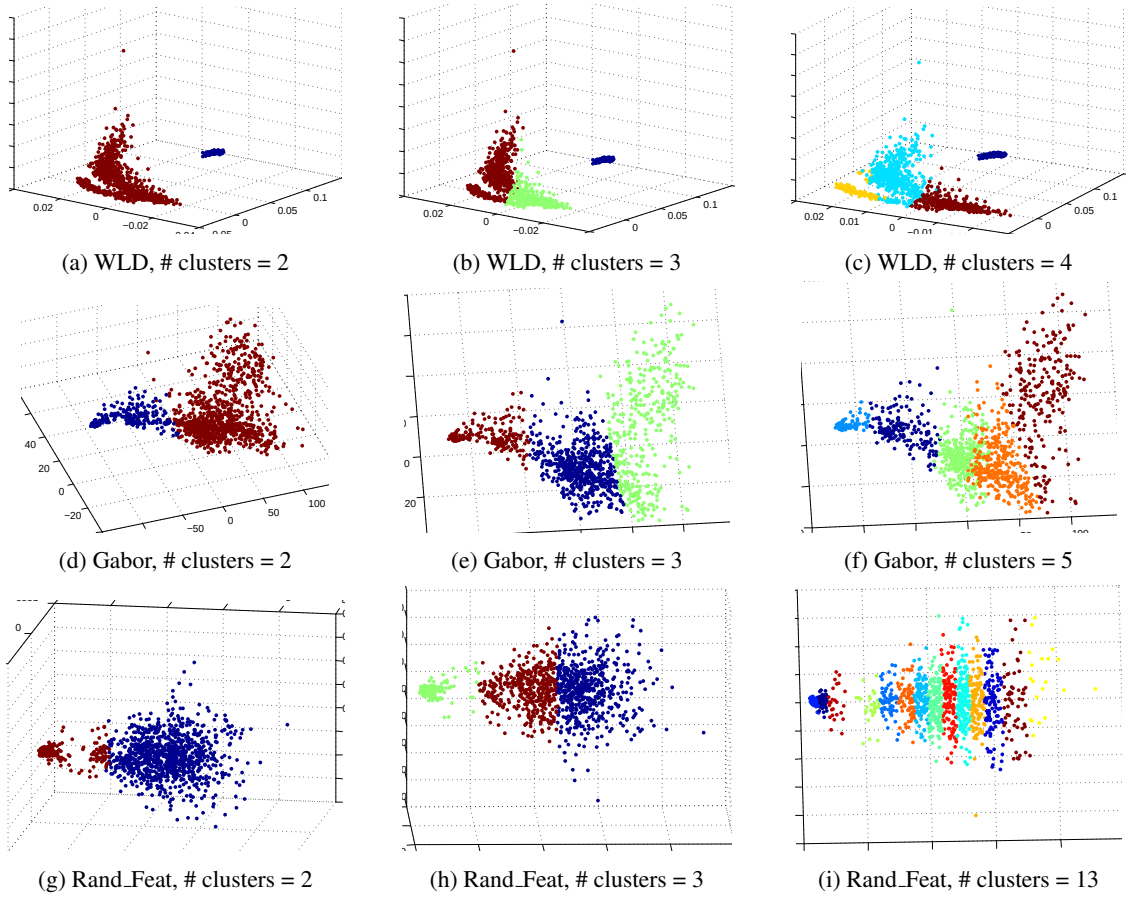
In this paper, we introduced a clustering-based approach to evaluate EO image indexing. In our proposed approach, the images are represented by feature descriptors. Then we perform clustering on the feature space, built by feature descriptors, for different number of clusters. The resulted clusterings are then evaluated both internally and externally (i.e., without and with using prior annotation, respectively). While internal evaluation demonstrate the structure of the feature space by finding the optimum clusters, external evaluation allows us to compare the image understanding by users and computers.
Experimental results show that the entire information provided by an image collection cannot be represented by only one kind of feature descriptors. Moreover, multiple optimum clusters can be found which corresponds to different levels of understanding of the images' contents (e.g., from low-level shape and texture to higher-level concepts).
External evaluation demonstrates that the structure discovered by computers (i.e, represented by feature descriptors) does not necessarily correspond to the prior annotation provided by users. Moreover, comparing the clusters for different feature descriptors shows that users consider some aspects of features in the images more than the others (e.g., consider textural features more than average intensity).

Evaluation of feature space provides users with better understanding of image collections. This allows the users to not only provide richer annotation, but also validate the annotation afterwards. Further, the study of feature space for different feature descriptors allows us to develop more sophisticated feature descriptors which not only group image collections to meaningful categories, but also provide relevant results to the users' queries.



(a) Average scattering (b) Average density between clusters (c) S_Dbw measure

Figure 3: Internal evaluation of clusters for three kinds of feature descriptors (e.g., WLD, Gabor, Rand_Feat) and different number of clusters.



(a) WLD, # clusters = 2 (b) WLD, # clusters = 3 (c) WLD, # clusters = 4

(d) Gabor, # clusters = 2 (e) Gabor, # clusters = 3 (f) Gabor, # clusters = 5

(g) Rand_Feat, # clusters = 2 (h) Rand_Feat, # clusters = 3 (i) Rand_Feat, # clusters = 13

Figure 4: 3D visualization of feature space for three kinds of feature descriptors (e.g., WLD, Gabor, Rand_Feat) including colored labeling for different clusters.

REFERENCES

Bahmanyar, R. and Datcu, M., 2013. Measuring the semantic gap based on a communication channel model. In: IEEE International Conference on Image Processing, ICIP 2013.

Caliński, T. and Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics - Theory and Methods 3(1), pp. 1–27.

Chen, J., Shan, S., He, C., Zhao, G., Pietikainen, M., Chen, X. and Gao, W., 2010. Wld: A robust local image descriptor. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32, pp. 1705 –1720.

Davies, D. L. and Bouldin, D. W., 1979. A cluster separation measure. Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-1(2), pp. 24–227.

FÃrber, I., GÃnnemann, S., Kriegel, H.-p., KrÃger, P., MÃller, E., Schubert, E., Seidl, T. and Zimek, A., 2010. On using class-labels in evaluation of clusterings.

Halkidi, M. and Vazirgiannis, M., 2001. Clustering validity assessment: finding the optimal partitioning of a data set. In: Data Mining, ICDM 2001, Proceedings IEEE International Conference on, pp. 187–194.

Hubert, L. and Arabie, P., 1985. Comparing partitions. Journal of Classification 2(1), pp. 193–218.

Larsen, B. and Aone, C., 1999. Fast and effective text mining using linear-time document clustering. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '99.

Liu, L. and Fieguth, P., 2012. Texture classification from random features. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(3), pp. 574–586.

Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J., 2010. Understanding of internal clustering validation measures. In: Proceedings of the 2010 IEEE International Conference on Data Mining, ICDM '10, IEEE Computer Society, Washington, DC, USA, pp. 911–916.

Manjunath, B. and Ma, W., 1996. Texture features for browsing and retrieval of image data. Pattern Analysis and Machine Intelligence, IEEE Transactions on 18(8), pp. 837–842.

Rendón, E., Abundez, I. M., Gutierrez, C., Zagal, S. D., Arizmendi, A., Quiroz, E. M. and Arzate, H. E., 2011. A comparison of internal and external cluster validation indexes. In: Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and applications, AMERICAN-MATH'11/CEA'11, World Scientific and Engineering Academy and Society (WSEAS), pp. 158–163.

Tan, P.-N., Steinbach, M. and Kumar, V., 2005. Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc.

Vinh, N. X. and Epps, J., 2009. A novel approach for automatic number of clusters detection in microarray data based on consensus clustering. 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE) 0, pp. 84–91.

Vinh, N. X., Epps, J. and Bailey, J., 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach. Learn. Res. 9999, pp. 2837–2854.