Evaluating Clustering Algorithms on a Multimedia Metadata Dataset

Christina S. Kyrkou
Department of Informatics
Ionian University
Corfu, Greece
chkyrkou@outlook.com

Phivos Mylonas
Department of Informatics
Ionian University
Corfu, Greece
fmylonas@ionio.gr

Evaggelos Spyrou
Institute of Informatics & Telecommunications
NCSR - "Demokritos"
Athens, Greece
espyrou@iit.demokritos.gr

Spyros Sioutas
Department of Informatics
Ionian University
Corfu, Greece
sioutas@ionio.gr

Angeliki Rapti
Computer Engineering and Informatics Department
University of Patras
Rio, Patras, Greece
arapti@ceid.upatras.gr

Dimitrios Tsolis
Department of Cultural Heritage Management & New Technologies
University of Patras
Rio, Patras, Greece
dtsolis@upatras.gr

ABSTRACT
In the framework of informatics, data analysis always played a crucial role in understanding various phenomena. When it comes to multimedia content, this task is getting more difficult to tackle in an efficient manner. In principle, cluster analysis, i.e., primitive data exploration with little or no prior knowledge, consists of research developed across a wide variety of approaches. The aim of this paper is to present a comparative survey of five clustering algorithms, namely k means, EM, DBSCAN, Mean Shift and KVQ applied on a real-life multimedia metadata dataset derived from Flickr social network.

Keywords
clustering; metadata; multimedia; empirical survey; geo-tagging

1. INTRODUCTION
Over the recent years the volume of digital still images produced and published on the Internet within social networks has grown rapidly, especially in virtual communities dealing with photo sharing. The reasons for the aforementioned observation lie on the expanding use of the Internet, the evolution of social behavior of its users, as well as the wide spread of digital cameras and smartphones. Nowadays, people may easily capture huge amounts of digital pictures in a short time spans. The latter phenomenon has created an urgent need for more efficient image search and retrieval.

The broad research field of image retrieval from large collections has set new challenges in a few other research fields, but mainly in those of machine learning and computer vision. Thus, many image retrieval algorithms have been developed. By using them, given an image or part of its metadata one may recover similar images in terms of high- or low-level features, or even their combination. Still, most image search algorithms rely on text-based approaches, e.g., they are based on keywords. In this framework great interest lies in the utilization of data mining techniques and particularly of clustering methodologies that can be applied to large image collections, for which there does not exist an adequate amount of information. These techniques may greatly improve both search and retrieval processes at any stage, such as in pre-processing, searching, visualization, storage etc.

The main goal of this work is to study, both at a theoretical and at a practical level utilizing common tools such as Weka [1] and Matlab [2], different methods that may be applied to manage metadata within such a multimedia information retrieval system. Specifically, in the context of this work, 5 well-known clustering algorithms (namely, k-means, EM, DBSCAN, Mean Shift, and KVQ) will be studied, examined and compared in terms of their efficiency, their suitability of use and according to their special characteristics. In the process of experimental validation, a photo metadata dataset derived from the VIRaL image retrieval system [3] shall be used. This set comprises of photos taken in downtown Athens, Greece, accompanied by their user-generated metadata. They have been collected from the popular social network Flickr[1]. The ultimate goal of this work is to assess whether the KVQ algorithm, which has been adopted in the context of the VIRaL system[2], is actually the most suitable for the needs of the particular application, as well as the extraction of generalized conclusions about selecting and using several clustering algorithms in different image retrieval applications by exploiting their metadata information.

The rest of this paper is structured as follows: in Section 2 we briefly present relevant research activities that acted as our motivation to research within this field. Section 3 discusses the main theoretical points of each clustering approach, as well as presents a comparative step-by-step analysis. In Section 4 we present our experimental results from the application of the five aforementioned algorithms to the utilized dataset. Finally, our respective conclusions are drawn in Section 5.

2. MOTIVATION AND RELEVANT WORK
The proliferation of multimedia content production and sharing within the Internet has led to the creation of huge,

[1] https://www.flickr.com/
[2] http://viral.image.ntua.gr/

PCI2016, November 10 – 12 2016, Patras, Greece.

increasing collections of photos. Combined information search and retrieval approaches attempt to provide solutions to the problems of organization, storage and search within this chaotic information that almost overwhelms the Internet. In order to achieve this, several approaches have been proposed over the last years. The dominant category of approaches is based solely on the textual metadata that accompany a photo. In other words, image retrieval problems are dealt as traditional text retrieval ones, using solely keywords in the process [4]. They allow for easy implementation, and enable quick retrieval; therefore, such approaches have been widely used on the Web for image search, by exploiting each image's surrounding textual metadata. However, since non-automatic annotation is common and due to the fact that metadata are not always available and accurate, the aforementioned approaches are prone to produce unreliable and inaccurate results. Still, together with DCNNs they are considered to be the most common approach currently used by the majority of popular image search engines on the Web, such as Google[3], Yahoo[4], and Bing[5]. As expected, search is conducted by considering solely the metadata describing the image, such as the title, the text and keywords. Over the recent years data-oriented content has also gained focus. More specifically, many private collections tend to incorporate metadata assigned to their multimedia content. As depicted in [5], metadata used to organize such collections can be classified according to several different ways, ranging from versatile to flat or no data at all.

Taking this a step further in most modern applications and research efforts the retrieval process utilizes a combination of both textual and visual information. In other words, textual and visual characteristics are provided, i.e., annotation or metadata as textual information, and low-level features (such as color, texture, etc.) as visual information. The idea behind this multimedia fusion is reaping the benefits of each method and the use of different sources as additional information to effectively complete a search and retrieval operation. In this manner multimedia fusion tries to aid solving the problem of the so-called “semantic gap”, while at the same time obtaining accurate results [6]. The same conclusion reaches also the work of Rui et al. [7], who argue that image retrieval based on content does not replace text-based retrieval, instead they tend to be complementary and their combination is required in order to obtain satisfactory retrieval results.

Now, as far as the grouping of objects and activities in accordance with some of their common features is concerned, this is a process that takes place, often intuitively, by humans on a daily basis, but usually goes unnoticed, precisely because it is rather automated. Cluster analysis or so called clustering is one of the most important techniques in data mining, in particular, and in computer science, in general. Strictly speaking, the term clustering refers to the clustering process of organizing a data collection into clusters based on a similarity measure [8].

Data is usually represented as measurements or vectors in a multidimensional space. From the above definition it should be clear that a cluster is a collection of objects that have similar behavior to each other and dissimilar with respect to objects of other groups. In research literature clustering often refers to as unsupervised learning; this is because it is not based on a priori defined classes and training samples with specific features per class, i.e., class-labeled examples. Quite on the contrary, data group themselves according to their similarities and differences. The fact that there is no prior knowledge about these groups separates clustering from classification, a technique that typically belongs to the supervised learning scheme.

Conference'10, Month 1–2, 2010, City, State, Country.

[3] https://images.google.com/
[4] https://images.search.yahoo.com/
[5] http://www.bing.com/images

3. THEORETICAL & COMPARATIVE ANALYSIS

3.1 Theoretical Analysis

The k-means algorithm is one of the simplest to use and most popular clustering algorithms for continuous numerical data. It belongs to the category of divisive methods and the original and most widely used approach has been formulated in [9]. Its main idea is to define *k* centers, one for each cluster, while *k* is a user-selected variable. We should emphasize the fact that, different locations of the initial centers would typically lead to different results. At the next step, every point of the dataset is assigned to its nearest center. When every point is associated with a center, a set of initial clusters has been formed. At this phase, the new centers are re-estimated, based on the created clusters. Subsequently, the distances between the points of the dataset and the new centers are recalculated and the points are reassigned again to the nearest center. This iterative process implies a greedy algorithm. Typically, k-means is assumed to converge to a local minimum, although recent studies indicate that it may converge to the global minimum with high probability, if the clusters are well-separated [10].

The Expectation-Maximization (EM) algorithm forms another important technique in data mining and at the same time is one of the most popular methods of likelihood maximization. Its fundamentals have been used by many experts in the past and in different variants; the best known approach is the one proposed by Dempster et al. [11], who introduced the term EM and demonstrated its convergence. EM forms an iterative process which can derive maximum likelihood estimates of parameters from observations, particularly in cases where there are - or may be assumed that there are - missing or hidden data.

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is a density-based clustering algorithm proposed by Ester et al. [12]. Its use is recommended in cases of clusters with a high density of points between them, which are separated from the points of lower density. It is successfully applied to 2D and 3D Euclidean spaces, as well as in some higher-dimension spaces. The main idea of the algorithm is that for each point of a cluster, its surrounding area (neighborhood) defined by a given radius, has to contain a minimum number of points; in other words, an area's density has to exceed a certain threshold.

The Mean Shift algorithm is a nonparametric clustering method based on density estimation introduced by Fukunaga et al. [13]. It is typically used for image analysis, low-level computer vision problems, etc. The mean shift approach is based on the idea of correlation of each point of the feature space with similar points. More specifically, the algorithm attempts to group a set of data in a previously unknown number of groups through the

identification of local maxima (modes) of a probability density function (pdf). In other words, the feature space can be considered as the empirical pdf of the represented parameter, and thus the dense areas in the feature space correspond to local maxima of the pdf, i.e., to the peaks of the unknown density.

The Kernel Vector Quantization (KVQ) algorithm proposed by Tipping et al. [14] is a kernel method that can be used for vector quantization and clustering, ensuring an upper bound of distortion and automatically adjusting the number of clusters based on it; the latter is one of the most important advantages of this algorithm. According to this method, the maximum distance between clusters may be considered as the maximum distortion level.

3.2 Comparative Analysis

From the above brief presentation, it is obvious that k-means is quite effective in terms of speed of execution and management of large datasets. On the contrary, it has the great disadvantage of the determination of the number of clusters by the user, as well as the proper initialization of the centers. The latter appears a huge problem, particularly in cases of large datasets, with no prior knowledge about their structure. However, it produces satisfying results in cases of clusters that are well separated and compact. Also, created clusters are typically of spherical (or elliptical) shape. Mean Shift and DBSCAN algorithms can produce arbitrary clusters, while EM, due to the probabilistic nature of the mixture distribution models, can produce arbitrarily structured clusters and not only spherical, through the selection of suitable density functions such as Poisson, the non-spherical Gaussian, etc. [15].

K-means and EM seem to have further similarities, since in general it has been observed that in datasets in which k-means provides good results, EM algorithm also provides good results. The same applies in cases of datasets in which k-means does not perform well, where the same thing happens with EM, which may be due to the selection of parameters [16]. In cases where it is needed to perform cluster analysis within small areas of interest and results produced by k-means are not satisfactory, a good alternative is the EM algorithm, which is based on a mathematical model and uses estimates for the parameters [17]. In contrast to k-means, EM is not based on distances, instead calculates possibilities for each observation to belong to a cluster, based on the selected distribution (which is usually the normal distribution). The ultimate goal of EM is to find the solutions (clusters) that maximize the overall data possibility.

Regarding the overlapping data points, k-means may not handle them satisfactorily. This happens, because it can only cluster a point based on its distance from the estimated centers. Thus, in case when data overlap, there is not a clear line that can be drawn to separate points that are closer to a center than those that are closer to another center. On the other hand, EM performs better with overlapping data. This happens because it has got the power to integrate basic assumptions about how the data had been initially created. The fact that the data overlap, is not of great importance, because the information needed by EM is the distance of the points from the centers of Gaussian kernels [16].

The next two algorithms, DBSCAN and Mean Shift, belong to the density-based category, i.e., they take into account the density of the points in order to form a logical number of clusters, although as it is already described they operate very differently. The DBSCAN algorithm even though it does not have the disadvantage of user-determined number of clusters, instead requires the determination of two parameters that affect to a great extend the actual results. The feature that differentiates DBSCAN from previously presented algorithms is its resilience to noise and the exclusion of noisy data points from any cluster. It is also suitable for large datasets. However, as every algorithm that uses the Euclidean distance as a measure of distance, it fails to provide satisfactory results in high-dimensional spaces and like other density-based algorithms, it is unable to cluster datasets with large variations in density [17]. The Mean Shift algorithm on the other hand is considered to be a non-parametric algorithm, and does not require the determination of any parameter related to the number or the shape of clusters; only an estimate about kernel bandwidth is needed instead. The base Mean Shift algorithm, i.e., when utilizing only Gaussian kernels, has been proved to be an EM algorithm. Contrary, when it uses a non- Gaussian kernel, is considered to be a generalized EM algorithm [17]. In practice, this means that it converges for almost any initial set of points, either by monotonically increasing the value of the density or by leaving it stable. Last but not least, KVQ is a kernel-based method that significantly differs from the rest of the aforementioned algorithms. It comprises a kernel approach that uses linear programming in order to find a small number of clusters which would cover the whole dataset. In the case of KVQ a point can belong to several clusters at the same time, so clusters may be overlapped. Unlike fuzzy algorithms and those that can be modified to operate as such, the points do not belong to each cluster with different weights or probability. Instead they may simultaneously belong to more than one, overlapping clusters. KVQ needs to calculate all distances between all the points and compare them with a threshold distance *r*. Then, it uses linear programming in order to find an appropriate set of centers (thus of clusters, too). The redundant centers are then removed in a pruning step. It operates completely different e.g., from k-means, who firstly sets the centers and then assigns a point to a single cluster based on the distance of the point from the temporary center of each cluster. It is also worth noting that unlike k-means and/or mean shift algorithms, the cluster centers are always points of the dataset and not just points of the vector space. Therefore, the new vectors that are produced are representative of the data. **Table I** summarizes comparative observations on the main characteristics of each algorithm.

TABLE I. CHARACTERISTICS OF THE STUDIED ALGORITHMS

Algorithm	Type of algorithm	Input parameters	Best for cases of…	Cluster shape	Noise man/ ment
k-means	partitioning	number of clusters, number of iterations	well separated clusters, big datasets, data with information about their structure	spherical, elliptic	no
EM	based on statistical model	initial estimates of gaussian parameters, convergence limit	big datasets with similar distribution	arbitrary and spherical	no
DBSCAN	density based	radius of clusters, minimum number of clusters in each cluster (eps, minpts)	big datasets, data with noise, data with small differences in density	arbitrary	yes

Mean Shift	density based	kernel bandwidth (h)	pursuit of semantically significant areas (e.g., image segmentation)	arbitrary	no
KVQ	kernel method	distortion parameter (r)	high dimensional spaces, overlapping clusters & semantically important, applications where upper distortion limit determination is an advantage (e.g., compression)	spherical	no

4. EXPERIMENTAL RESULTS

In this Section we shall provide a presentation of geographic clustering results produced by the application of the aforementioned algorithms on a real-life geo-tagged dataset. Towards this goal we used Weka v.3.6 implementations of k-means, EM and DBSCAN, whereas Matlab v.R2015a has been used for Mean Shift and KVQ. Then, we conducted experiments with different sets of parameters, in order to be able to provide a comparative study. Next, we will describe in detail the selection of the appropriate parameters, the application of the algorithms and the visualization of the produced results.

The dataset used has been derived from the VIRaL system and consists of 2500 pictures with geographical annotations (i.e., geo-tagged photos) from downtown Athens, Greece, along with their accompanying metadata. Each photograph was in .jpg format and was accompanied by a relative text document (.txt file)[6] containing the actual metadata. It should be noted at this point that herein we focused on the use of metadata only and not on the visual content of the images. As a result photographs were used only for human evaluation and verification of the data content.

Our methodology was based on the assumption that images captured in a nearby area are more likely to reflect the same subject [3]. As a result, from the available metadata we only considered the geographical coordinates, i.e., the latitude and longitude, where photos had been taken (strictly speaking, these coordinates are the ones automatically added by the camera or manually added by the photographer, which in some cases introduces errors) and we shall compare the five algorithms based on these coordinates. For the sake of completeness, we also maintained the PhotoId feature, which is the unique ID of each image capture within Flickr. We depict a sample of the initial distribution of the geographical information of the dataset around the Acropolis landmark in **Figure 2**. In the case of **k-means** the seed number is used to initialize a random number generator. Different clusters are formed each time due to the different initialization of the centers, whereas best results are obtained for 30 clusters with seed = 12 (see **Table II**, **Figure 1** and **Figure 2**). We observe that in this case all points have to be included in a cluster, meaning that even points which are far from the center of any cluster are required to join a particular cluster (**Figure 3**); an outcome not always wished in geographical clustering.

TABLE II. K-MEANS RESULTS, SEED = 12

Seed 12			
# clusters	# iterations up to convergence	square error	execution time (in secs)
60	16	0.7008186107525284	0,41
55	18	0.7027029997704914	0,48
50	14	0.7679658848787074	0,28
45	22	0.7836945818882881	0,42
40	19	0.9808141569408151	0,31
35	38	0.9844519096065154	0,56
30	35	0.9326740914219306	0,47
25	36	0.9645608297932126	0,39
20	30	1.4859384876779858	0,28
15	20	2.5523634192078495	0,14

Figure 1. K-means results for 30 clusters and a seed of 12.

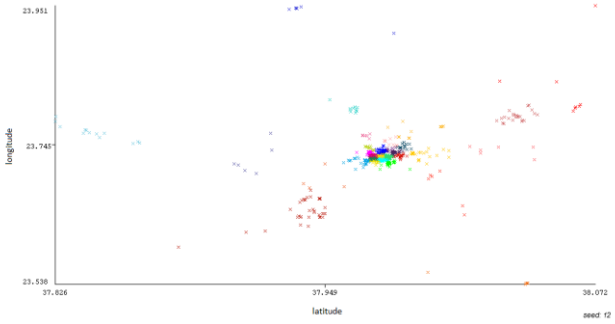


Figure 2. K-means; instance ranking for seed 12.

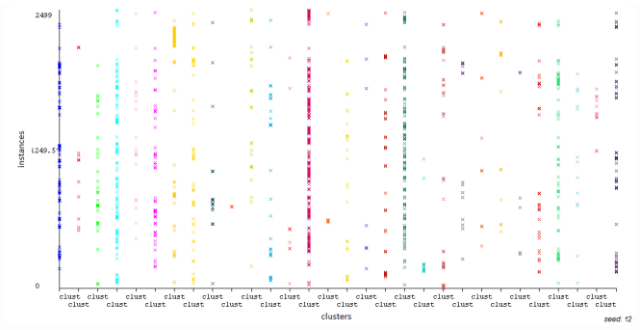
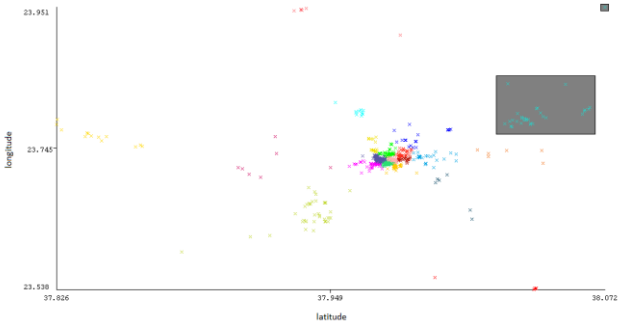


Figure 3. K-means; remote point integration example.



As depicted in **Table III**, the goal of **EM algorithm** is to maximize the value of the logarithmic likelihood, therefore, the larger its values are, the better results we obtain. According to our experiments, better results are obtained for the case of 50 clusters with a seed value = 100, which gives us the maximum logarithmic likelihood value (i.e., 9.01575).

TABLE III. EM RESULTS, SEED = 100

Seed 100			
# clusters	# iterations	logarithmic likelihood	Execution time (in secs)
-1 -> 11	10	7.13786	80.08
-1 -> 13	100	7.802	156.12
-1 -> 13	500	7.802	155.64
-1 -> 13	1000	7.802	156.53
30	28	8.09785	3.57
35	35	8.23716	4.31
40	27	8.30859	4.87
20	100	8.03668	2.59

[6] http://www.image.ntua.gr/~fmylonas/viral-sample/

30	100	8.09785	3.53
35	100	8.23716	4.41
40	100	8.30859	4.94
45	100	8.60277	5.34
50	100	9.01575	5.84
55	100	7.7558	4.85
60	100	8.06558	5.52

Regarding execution times, they tend to be generally increasing mainly because of the number of repetitions, and in any case are greater than the ones required by k-means. For instance, considering the case of 30, 35 and 40 clusters (with seed = 100), where for the same number of repetitions needed by k-means to converge, EM delivers longer times. Similarly to k-means, distant points are often clustered together (**Figure 4** & **Figure 5**). In addition, EM may form clusters of arbitrary shape, even provide a cluster within another cluster, which in principle is a property not desirable in the case of geographical data analysis (**Figure 6**).

Figure 4. EM results, 50 clusters, seed = 100, 100 iterations

Figure 5. EM - Example of integration of remote points in 1 cluster

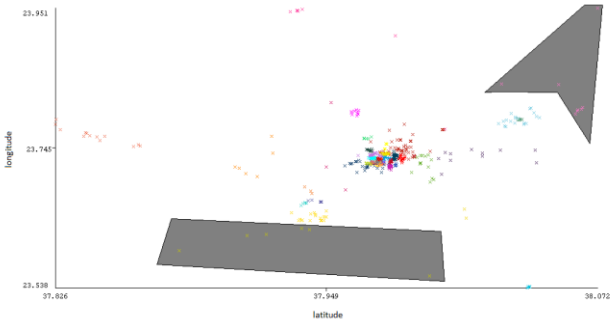
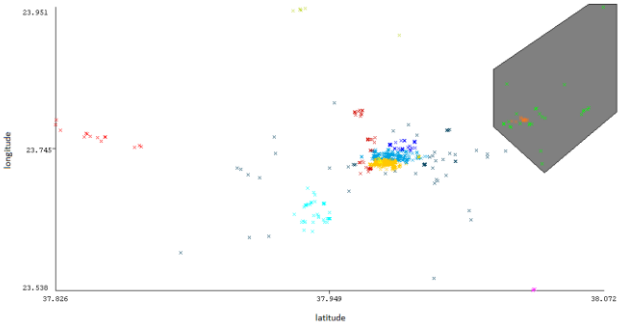


Figure 6. EM - Example of finding a cluster enclosed by another, for 100 iterations, and automatic estimation of the number of clusters



Selection of parameters in **DBSCAN** was based on the actual dataset available; iterations were conducted using MinPts parameter equal to 4, 5 and 6. Furthermore, EPS radius, which actually defines the minimum size of a cluster and is a measure of the average distance of points, was selected to range between 0.002 and 0.05. It should be noted that these somehow low values of the parameter are fully justified considering the fact that points represent photographs taken exclusively in the Athens downtown area, and therefore they are typically conveniently located very close to each other. **Table IV**, **Figure 7** and **Figure 8** illustrate the according results.

Figure 7. DBSCAN results for Eps = 0,03 and MinPts 5

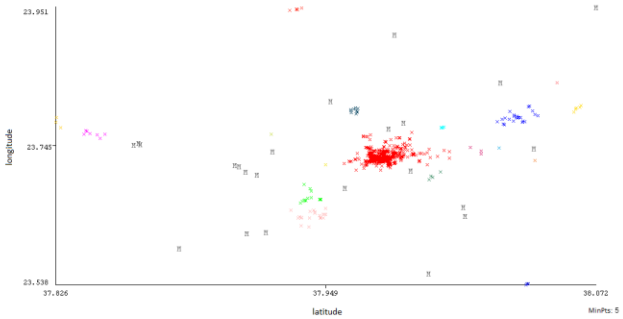


Figure 8. DBSCAN results for Eps = 0,008 and MinPts 6

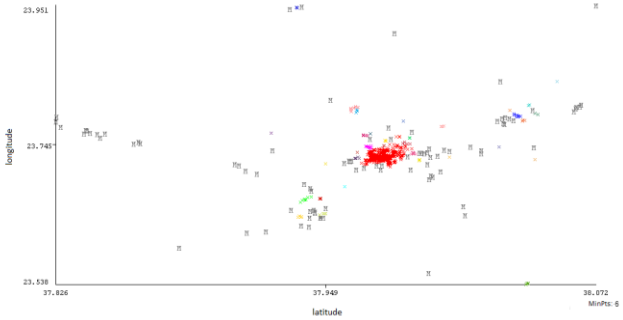


TABLE IV. DBSCAN RESULTS FOR MINPTS 4-6, EPS 0,002-0,05

MinPts	Eps	# clusters	# noise points	execution time (in secs)
4	0.002	76	285	1.05
4	0.004	51	183	1.31
4	0.006	44	127	1.35
4	0.008	35	112	1.31
4	0.010	32	90	1.39
4	0.030	20	29	1.52
4	0.050	15	16	1.50
5	0.002	56	369	1.14
5	0.004	40	227	1.35
5	0.006	34	167	1.34
5	0.008	27	145	1.30
5	0.010	24	127	1.35
5	0.030	18	37	1.64
5	0.050	12	28	1.43
6	0.002	53	409	1.14
6	0.004	33	262	1.26
6	0.006	26	211	1.27
6	0.008	23	165	1.30
6	0.010	19	154	1.41
6	0.030	10	77	1.73
6	0.050	8	48	1.96

The implementation of the **mean shift** algorithm uses a flat kernel and was executed for bandwidth values between 0.001 and 0.05. Greater bandwidth causes merging of clusters, as it is the case with the increase of Eps in DBSCAN. As the bandwidth decreases, the number of clusters increases, but as shown graphically (see Figure 11), points which are very close to other points, and should intuitively belong to the same cluster, because of the density difference, tend to create a single cluster, resulting in many small meaningless clusters. Also, in general we may observe that as long as the bandwidth is reduced, execution time increases. **Table V** and **Figure 9** illustrate the Mean Shift results.

Figure 9. Indicative Mean Shift results for bandwidth value 0.03; black dots represent the centers of the 116 clusters.

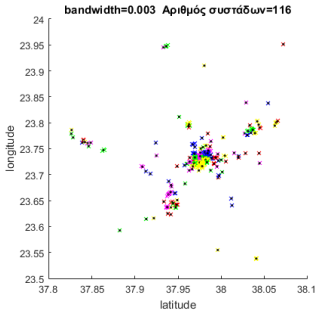


TABLE V. MEAN SHIFT RESULTS FOR BANDWIDTH 0.001-0.05

Bandwidth	# clusters	execution time (in secs)
0.001	249	0.285412
0.002	156	0.166827
0.003	116	0.152521
0.004	92	0.132602
0.005	80	0.074463
0.006	66	0.077843
0.007	59	0.072986
0.008	51	0.055245
0.009	43	0.055093
0.010	41	0.069771
0.020	21	0.034915
0.030	16	0.019610
0.040	10	0.014742
0.050	8	0.017690

Finally, **KVQ** generates overlapping clusters, centered on some dataset points. In general clusters appear to be more representative than before. The main benefit of the algorithm is that individual points being away from any other point, create clusters of their own and are not necessarily assigned at clusters that are not related to. In addition, a pruning step significantly reduces the number of clusters that would otherwise be meaningless. Despite its advantages, we should also note here that the step of calculating all distances between the points requires major computing resources and leads to considerably larger runtimes compared to all previous methods. **Table VI** and **Figure 10** illustrate KVQ results.

TABLE VI. RESULTS OF KVQ ALGORITHM

parameter r	# clusters before pruning step	# clusters after pruning step	execution time incl. pruning step (in secs)
0.005	295	87	141.472487
0.006	295	78	169.930471
0.007	261	69	196.236822
0.008	192	55	151.758631
0.010	209	48	326.494207
0.020	198	26	310.208670
0.030	157	18	659.986884
0.040	79	15	1074.366037
0.050	70	13	722.303342

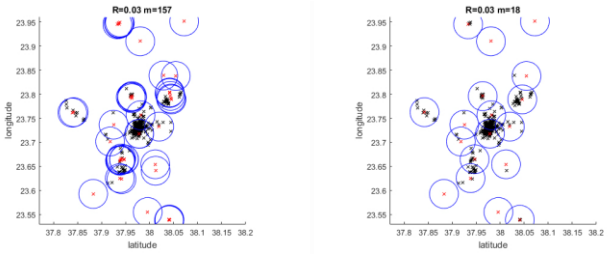
5. CONCLUSION

In this work we presented an evaluation study of 5 popular clustering algorithmic approaches applied on a multimedia metadata dataset derived from VIRaL system. By utilizing tools like Matlab and Weka in the process, we experimentally compared typical algorithms of related clustering literature, namely: k-means, EM, DBSCAN, Mean Shift, and KVQ. According to our qualitative findings against the ground truth of the dataset, the last kernel-based method produces the best results, although some limitations, such as high memory requirements and its actual execution time, hinder the overall process.

6. ACKNOWLEDGEMENT

Our thanks to C.Caratheodory Research Program from University of Patras, Greece to support this research.

Figure 10. Indicative KVQ results for r equals to 0.03; left/right column shows clusters before(157)/after(18) the pruning step; red x depicts a cendroid, blue circle the area of a cluster with radius r.



7. REFERENCES

[1] http://www.cs.waikato.ac.nz/ml/weka/, last retrieved on 30/05/2016

[2] http://www.mathworks.com/products/matlab/?requestedDomain=www.mathworks.com, last retrieved on 30/05/2016

[3] Y. Kalantidis, G. Tolias, Y. Avrithis, M. Phinikettos, E. Spyrou, Ph. Mylonas, S. Kollias, VIRaL: Visual Image Retrieval and Localization. Multimedia Tools Appl. 51, 2, January 2011, 555-592.

[4] G. Ahmed, R. Barskar, "A study on different image retrieval techniques in image processing.", International Journal of Soft Computing & Engineering (IJSCE), vol. 1, no. 4, pp.247-251, 2011.

[5] K. Yee, K. Swearingen, K. Li and M. Hearst, "Faceted metadata for image search and browsing", in SIGCHI Conference on Human Factors in Computing Systems, Florida, 2003, pp. 401-408.

[6] V. Khobragade, L. Patil and U. Patel, "Image retrieval by information fusion of multimedia resources", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 4, no. 5, pp. 1721-1727, 2015.

[7] Y. Rui, T. Huang, S. Chang, "Image Retrieval: Current Techniques, Promising Directions, & Open Issues", Journal of Visual Comm. & Image Representation, vol. 10, no. 1, pp. 39-62, 1999.

[8] C. Manning, P. Raghavan and H. Schütze, Introduction to information retrieval. New York: Cambridge University Press, 2008.

[9] J. MacQueen, "Some methods for classification and analysis of multivariate observations.", in Proc. of 5th Berkeley symposium on mathematical statistics & probability, 1967, pp. 281-297.

[10] A. Jain, "Data clustering: 50 years beyond K-means", Pattern Recognition Letters, vol. 31, no. 8, pp. 651-666, 2010.

[11] A. Dempster, N. Laird and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society.", Journal of the royal statistical society. Series B (methodological), vol. 39, no. 1, pp. 1-38, 1977.

[12] M. Ester, H. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise.", in KDD - 9 6 Proceedings AAAI, 1996, pp. 226-231.

[13] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition", IEEE Trans. Inform. Theory, vol. 21, no. 1, pp. 32-40, 1975.

[14] M. Tipping and B. Schölkopf, "A kernel approach for vector quantization with guaranteed distortion bounds.", Artificial Intelligence and Statistics, pp. 129-134, 2001.

[15] P. Bradley, U. Fayyad, C. Reina, "Scaling EM clustering to large databases", Microsoft Research, 1998.

[16] N. Bhan, D. Mehrotra, "Comparative study of EM & k means clustering techniques in weka interface", International J. of Advanced Tech. & Engineering Research, 3(4), pp. 40-44, 2013.

[17] N. Sharma, A. Bajpai and R. Litoriya, "Comparison the various clustering algorithms of weka tools", International J. of Emerging Technology & Advanced Engineering, 2(5), pp.73-80, 2012.

[18] M. Carreira-Perpinan, "Gaussian Mean-Shift Is an EM Algorithm", IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(5), pp. 767-776, 2007