

# A comparison of clustering methods for word image indexing

Simone Marinai, Emanuele Marino, and Giovanni Soda  
Dipartimento di Sistemi e Informatica - Università di Firenze,  
Via S.Marta, 3 - 50139 Firenze - Italy  
marinai@dsi.unifi.it

## Abstract

*In this paper we explore the effectiveness of three clustering methods used to perform word image indexing. The three methods are: the Self-Organizing Map (SOM), the Growing Hierarchical Self-Organizing Map (GHSOM), and the Spectral Clustering. We test these methods on a real data set composed of word images extrapolated from pages that are part of an encyclopedia of the XIX<sup>th</sup> Century. In essence, the word images are stored into the clusters defined by the clustering methods and subsequently retrieved by identifying the closest cluster to a query word. The accuracy of the methods is compared considering the performance of our word retrieval algorithm developed in our previous work. From the experimental results we may conclude that methods designed to automatically determine the number and the structure of clusters, such as GHSOM, are particularly suitable in the context represented by our data set.*

## 1. Introduction

Digital Libraries store very large collections of documents in image format. From the user point of view, it is always possible to access the images by browsing the table of contents, but it is more complex to perform word searches in the free text, as we usually make with Internet search engines, such as Google.

The retrieval of images by text content is possible with techniques derived from Document Image Retrieval (DIR). DIR aims at finding relevant document images from a corpus of digitized pages and it is a research field that lies at the borderline between classic Information Retrieval (IR) [1] and Content Based Image Retrieval (CBIR) [7].

In this paper, we focus on one specific sub-task of DIR: the retrieval of documents on the basis of the textual content. We describe an approach to perform

the efficient identification of the occurrences of a given word in the indexed documents. This task had already been studied in a previous work [4], where we introduced a word indexing general system designed to manage large collection of document images. In [4] we proposed a SOM-based word image clustering to speed up the word retrieval.

In this paper, we explore the effectiveness of two clustering algorithms in comparison with standard SOM: the Growing Hierarchical Self-Organizing Map (GHSOM) [2, 9] (that is a SOM-based method) and Spectral Clustering [6, 5]. In particular, in this paper we consider the Spectral Clustering algorithm proposed in [6], that allows the automatic determination of the number of clusters.

The present work is organized as follows: in Section 2 we briefly describe the clustering methods that will be object of our analysis. Section 3 is dedicated to the description of the word indexing and retrieval approach. In Section 4 some experimental results are reported and finally the conclusions are drawn in Section 5.

## 2 Word Images Clustering

The proposed word indexing is composed by several steps [4], as summarized in Section 3. In this paper we focus on the clusters determination. In [4] we used a SOM-based approach, but in recent years novel clustering techniques emerged, and in particular the number of algorithms that allow the automatic determination of clusters. These algorithms seem to be very interesting in our context, where it is difficult to estimate the number of clusters a priori.

### 2.1 Spectral Clustering

Spectral clustering involves constructing an affinity matrix from the data and requires, in the original version, the prior knowledge of the number of clusters.

The only parameter that we have to set is the variance value  $\sigma^2$ , necessary to the construction of the affinity matrix.

In this paper we consider the modified algorithm proposed in [6], which allows to iteratively obtain the number of clusters.

Given a dataset consisting of  $N$  data vectors  $x \in R^d$ , the first step is to construct the affinity matrix  $A \in R^{N \times N}$  defined by

$$A_{ij} = \exp\left(\frac{-\|x_i - x_j\|^2}{\sigma^2}\right) \quad (1)$$

which is normalized obtaining the matrix  $L$ :

$$L = D^{-1/2} A D^{-1/2} \quad (2)$$

where  $D = \text{diag}(\sum_{j=1}^N A_{ij})$ . Compute the  $K$  largest eigenvectors of  $L$  and arrange them in an  $N \times K$  matrix  $Y$ . At this point, the original data set is projected in a  $K - \text{dimensional}$  vector space, and performing K-means (or other clustering algorithm) on these vectors will return the desired clustering. The Algorithm proposed in [6] is based on k-means and starting from  $K = 2$ , it increments iteratively  $K$  till no more extra-clusters is detected.

The software used in our experiments is the Spectral Clustering Matlab package proposed in [6].

## 2.2 SOM

The Self-Organizing Map (SOM) is a type of artificial neural network that is based on unsupervised learning [3]. In the SOM the neurons are typically arranged in a two dimensional lattice. Each neuron of the SOM map is associated with a weight vector  $w_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T \in \mathfrak{R}^n$ . During learning, all the input vectors  $x \in \mathfrak{R}^n$  are compared with all the weight vectors of the SOM map and the Best Match Unit (BMU) is determined: the BMU is the weight vector closest to the input vector according to a pre-determined distance, for example the Euclidean distance. The BMU  $c$  is therefore defined by

$$c = \arg \min_i \|x - w_i\| \quad (3)$$

where  $x$  is an input vector and  $w_i$  is the weight vector associated with neuron  $i$ , named *centroid* or *code-book vector* in SOM's terminology. After the BMU is determined, the input vector  $x$  is mapped into the cluster represented by neuron  $c$ , and this neuron and its neighbors are updated. In the update process, two types of rules may be used: the *incremental* learning rule and the *batch* learning rule.

In the incremental learning the update process is performed using each data vector at a time, and the centroid is updated by rule defined as:

$$m_i(t+1) = m_i(t) + h_{ci}[x(t) - m_i(t)] \quad (4)$$

where  $t$  denotes time,  $x(t)$  is the input vector,  $m_i(t)$  is the centroid to update and  $h_{ci}$  is the neighborhood function [3]. In batch training algorithm, at each step the whole data set is partitioned along the neurons of the SOM map without performing any update. After partitioning the data set, all neurons are updated according to the following rule:

$$m_i(t+1) = \frac{\sum_{j=1}^n h_{ci} x_j}{\sum_{j=1}^n h_{ci}} \quad (5)$$

At the end of each training step, the new centroid  $m_i(t+1)$  is therefore a weighted average of the data vectors belonging to the  $i - th$  cluster.

The SOM software used in our experiments is the SOM Matlab Toolbox [8], that is more recent than SOM\_PAK [3] used in [4]. The SOM Matlab Toolbox allows to choose between incremental or batch training, contrarily to SOM\_PAK where only incremental algorithm is implemented.

## 2.3 GHSOM

The need to predefine the SOM structure results in a significant limitation on the final mapping achievable [9]. To avoid the problem of the static SOM structure, several algorithms based on the basic SOM have been recently proposed. The Growing Hierarchical Self-Organizing Map (GHSOM) consists of a method of dynamically modelling the data set that is presented [2]. The GHSOM allows the network structure to grow in two dimensions, in width and in depth, so that it combines the advantages of the dynamic growth and the hierarchical structure. The growing process of the GHSOM is regulated by two parameters:  $\tau_1$  and  $\tau_2$ . The parameter  $\tau_1$  serves to regulate the growing process in width while  $\tau_2$  the growing process in depth.

Each layer of the trained GHSOM consists in several independent SOMs. The map at layer 0 consists of a single-unit SOM and serves as a representation of the complete data set. This single-unit is a weight vector  $m_0 = [\mu_{01}, \mu_{02}, \dots, \mu_{0n}]^T$  and is computed as the average of all input data.

The GHSOM training starts with its first SOM layer, consisting of a grid of  $2 \times 2$  neurons (his parent is the single-unit SOM at layer 0). After a fixed number  $\lambda$  of training iterations, the unit  $E$  with the highest mean

quantisation error is identified, and a row or a column of neurons is inserted between the unit  $E$  and its most dissimilar neighboring unit (the neighboring unit with the largest distance from the unit  $E$  in terms of input space). After this insertion, the training process is reset to its initial values and the training restart with the standard SOM algorithm. The growing process in width is repeated as long as  $MMQE_m \geq \tau_1 \cdot MMQE_p$  is true for map  $m$ , where  $p$  is the parent map of  $m$  and the mean quantisation error of the map  $m$  is defined by:

$$MMQE_m = \frac{1}{n} \cdot \sum_{i=1}^n MQE_i \quad (6)$$

where  $n$  is the total number of neurons of the map  $m$  and  $MQE_i$  is the quantisation error of the unit  $i$ . When the training of a map  $m$  is finished, every unit  $i$  that satisfies a predetermined condition is subjected to hierarchical expansion: this condition is  $MQE_i > \tau_2 \cdot MMQE_0$ , where  $MMQE_0$  is the mean quantisation error over the single-unit SOM at layer 0.

The package used in our experiments is the GHSOM Toolbox for Matlab. This toolbox is developed by authors in [2].

### 3 Word Indexing and Retrieval

The word image indexing system that we proposed in [4] is composed by the following steps: 1) word segmentation 2) SOM training with a sub-set of the words to be indexed. Principal Component Analysis (PCA) computation from the training patterns in each cluster. 3) Projection of all the indexed words in a separated lower dimensional space for each cluster.

In this paper we explore the effectiveness of other clustering methods in comparison to the standard SOM. The three clustering methods examined have been tested taking into account step 2 of our word indexing system. Some changes have been required in order to adapt GHSOM and Spectral Clustering to our indexing and retrieval algorithms.

#### 3.1 Indexing

After performing segmentation, each word image is processed obtaining one simple feature vector that is based on word image zoning. This standard technique consists of overlapping the word image with a fixed-size grid and computing some features in each grid cell. For each grid cell the average gray level of the pixels is computed, obtaining an high dimensional feature vector  $x \in R^n$  (hundreds of dimension): this word coding

technique allow us to cast the word retrieval as a problem of search in high dimensional vector space.

At this point, the system is ready to start training in order to determine clusters. After training, the whole data set is stored into the clusters. Next, the projection that best represents the data is computed by PCA. As mentioned in [4], our SOM-based word indexing system allows to efficiently address the word retrieval problem, and the reason is that during retrieval it is not needed to scan all the indexed words. Moreover, the PCA computation allows us to perform an efficient search in the projected space for each cluster. In Fig. 1 an example of two clusters is showed. The cluster on the left is more confused, and this fact is reflected on centroid computation during training. The cluster on the right contains homogenous data, and therefore its centroid looks like a real word image. The words in the two clusters are sorted by means of the distance from centroid.



Figure 1. Two examples of clusters and their centroid are showed.

To adapt GHSOM to our system, we had to decide in which clusters of GHSOM structure word images will be stored. After training, GHSOM presents a tree structure, where the nodes are the generated SOMs. For each independent SOM, the clusters suitable to store words are only those clusters that are not subjected to hierarchical expansion during training. If we consider the top part of Fig. 2 (where is reported a two layer GHSOM), these clusters are identified with gray color in the principal grid (which is a first layer

SOM), and all the remaining clusters belonging to all SOMs of the second layer.

### 3.2 Retrieval

During word retrieval, we first identify the  $K$  clusters having the centroid closer to the query word image (generated by L<sup>A</sup>T<sub>E</sub>X starting from an user-defined text query), and then we look for the  $q$  nearest neighborhoods inside each of the  $K$  clusters, by means of PCA. The PCA projection approximates the similarity in the  $n$  dimensional space, therefore we require a refinement step to obtain the final rank. The final rank is obtained computing the euclidean distance in the original space between the query word and the words belonging to the list obtained merging the  $K$  word lists.

To adapt Spectral Clustering method to our system, we had to compute the centroids for the generated clusters. The centroids have been represented simply by the median of patterns stored in each cluster. The median has been computed after training.

## 4 Experiments

The three clustering methods are tested on a real data set consisting of 132,956 word images extrapolated from 1,302 pages that are part of an encyclopedia of the XIX<sup>th</sup> Century<sup>1</sup>. We used the same training set for all methods that consists of a sub-set containing 6,650 word images. After training, the entire data set has been stored in the clusters generated by the three methods.

The SOM trained had a grid structure of 17×24 neurons, so we had a total of 408 clusters. The batch training algorithm is used and the dimension of the map was automatically determined by heuristic function adopted in SOM toolbox. The GHSOM has been trained by setting the parameters  $\tau_1 = 0.6$  and  $\tau_2 = 0.005$ , and a two-layer map has been generated and total number of clusters was 1980. Each independent SOM of GHSOM structure has been trained with batch training algorithm. The Spectral Clustering algorithm has been ran by setting the parameter  $\sigma^2 = 1000$ , and 180 clusters have been generated. The value to retain for parameters has been obtained after some preliminary tests. In Fig. 2 the GHSOM trained with our train set is reported.

The effectiveness of the clustering methods has been measured in term of accuracy achieved by our word retrieval algorithm [4] performing several queries and

computing a Precision-Recall plot [1] for each query. During searching of the query word into the clusters a  $q = 20$  nearest neighborhoods was performed.

The number of words to be used as queries was 70 and they have been selected taking into account rare words and frequent words: the set of query words was composed for example of 20 words with a number of occurrences between 1 and 3, 9 words in the range between 40 and 80 occurrences, and 8 words with more than 100 occurrences (the top words were *France* with 576 occurrences, *produit* with 571, and *Europe* with 273).

We compared the three methods with Sequential Scan, that consists on the linear comparison of each indexed word with the query word. We conclude that the efficiency gain is not obtained at the cost of a reduced retrieval effectiveness. In order to show an homogenous comparison between the three clustering methods, during test we ran our retrieval algorithm by setting the  $K$  parameter considering the percentage of clusters generated. We considered the following percentage values: 1, 5, 10 and 15 percent. In Fig. 3 the plot average Precision at 0 Percent Recall [1] vs the percentage of clusters is reported. In the plot the percentage of clusters is obviously reported only for the clustering methods. From the plot of Fig. 3 we observe that in correspondence of 5 percent of clusters, the three clustering methods achieve an accuracy that is approximatively the best accuracy achievable. From this plot, it is also possible to verify that GHSOM achieves best performance in term of accuracy in comparison with Sequential Scan, and this result is obtained with a gain in term of efficiency if we consider the percentage of clusters involved during the test. From the plot of Fig. 3 we also conclude that Spectral Clustering is not suitable in our real data set: it may depends on performance of the k-means or on distribution of the word encoding vectors of data set involved in the clustering algorithm.

In Fig. 4 we report a plot representing the percentage of patterns of data set involved during test in correspondence of number of clusters. From this plot we observe that all three methods process less than 10 percent of patterns in correspondence of 5 percent of clusters. This analysis is important to investigate the efficiency issue in comparison with the Sequential Scan.

From the analysis of the two plots, we can note some interesting aspects regarding the comparison between GHSOM and SOM. We can observe that in correspondence of 1 percent of the clusters, GHSOM achieves best accuracy but the two methods approximatively process the same number of the data vectors (GHSOM 2.33 percent of patterns, while SOM 1.18 percent of patterns). On the other side, when considering 10 per-

<sup>1</sup>*Les Merveilles De l'Industrie*: downloaded from the web site of the *National Library of France* (<http://gallica.bnf.fr>).

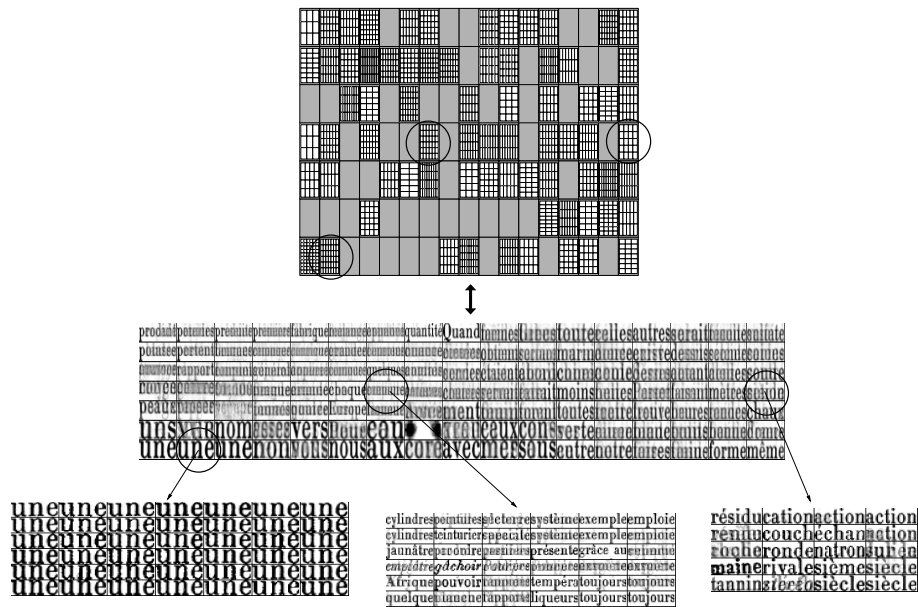


Figure 2. Top: the two-layer GHSOM is showed. The root map has a grid structure of  $7 \times 17$  neurons and the neurons not expanded in depth are identified in gray color. Bottom: the same structure is showed but we report all the image centroids of the root map and the image centroids of some maps belonging to second layer. Clusters highlighted with a black circle belonging to the map of the top part correspond to the highlighted clusters belonging to the map of the bottom part.

cent of clusters the gap in term of accuracy is reduced, but SOM processes 12.82 percent of patterns while GHSOM 16.86 percent: so, we have a significant gain in term of efficiency by SOM.

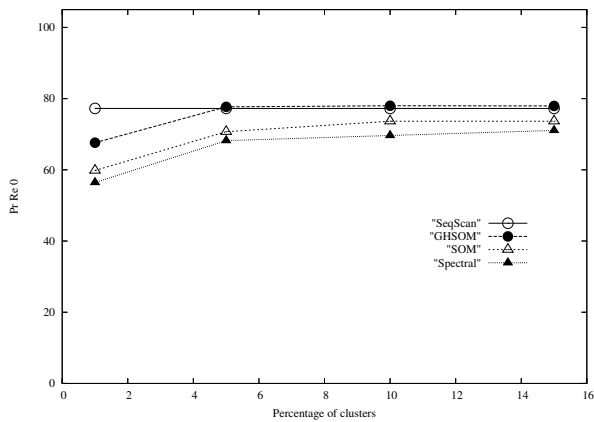


Figure 3. Precision at 0 Percent Recall vs Percentage of Clusters.

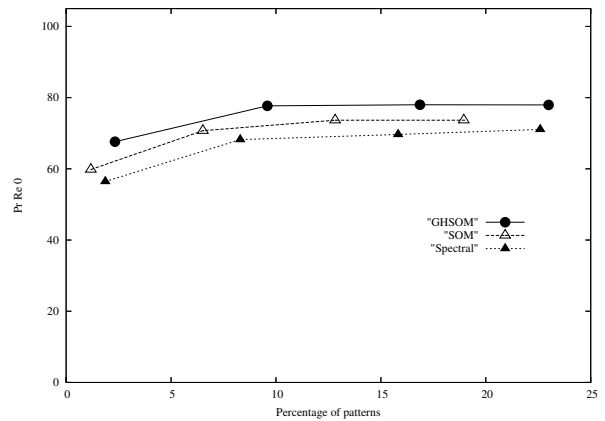
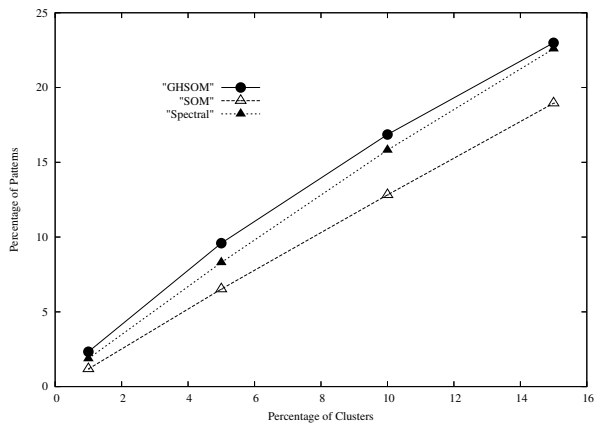


Figure 4. Precision at 0 Percent Recall vs Percentage of Patterns.



**Figure 5. Percentage of Patterns vs Percentage of Clusters.**

## 5 Conclusions

We explored the effectiveness of three clustering methods, the Self-Organizing Map (SOM), the Growing Hierarchical Self-Organizing Map (GHSOM) and Spectral Clustering. These methods are tested on a real data set consisting of 132,956 word images. We conclude that GHSOM is a very promising method to address the issue of word image indexing in Digital libraries, and more in general we conclude that to achieve good performance in term of accuracy, it is needed to develop a system based on a clustering method of dynamically modelling the data set.

Some aspects will require additional investigations: the introduction in training and indexing of more appropriate word images distances instead of Euclidean distance, that is the default distance adopted in all three methods analysed in this paper, and the use of other clustering algorithms in Spectral Clustering method in opposition of k-means.

## References

- [1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
- [2] E. Chan, A. Pampalk. Growing hierarchical self organising map (ghsom) toolbox: visualisations and enhancements. *Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference on*, 5:2537–2541, November 2002.
- [3] T. Kohonen. *Self-Organizing Maps*. Springer Series in Information Sciences, 2001.
- [4] S. Marinai, S. Faini, E. Marino, and G. Soda. Efficient word retrieval by means of som clustering and pca. *7th International Workshop on Document Analysis Systems, Nelson (New Zealand)*, 2006.
- [5] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems 14: Proceedings of the 2001*, 2001.
- [6] J. Sanguinetti, G. Laidler and N. D. Lawrence. Automatic determination of the number of clusters using spectral algorithms. *Machine Learning for Signal Processing, 2005 IEEE Workshop on*, pages 55–60, September 2005.
- [7] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1349–1380, 2000.
- [8] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas. Som toolbox for matlab. *J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas. SOM Toolbox for Matlab 5. Technical Report A57, Helsinki University of Technology*, <http://www.cis.hut.fi/projects/somtoolbox/>, April 2000.
- [9] G. Z. W. Yen. Ranked centroid projection: A data visualization approach with self-organizing maps. *Neural Networks, IEEE Transactions on*, 19:245–259, February 2008.