

# Using Sequential Clustering for Unsupervised Classification of Unconstrained Document Images

D. Curtis, J. Tse, C. Jones, E.A. Yfantis, S. Miller

Digital Image Processing Laboratory  
Department of Computer Science  
University of Nevada, Las Vegas  
Las Vegas, Nevada, USA

**Abstract** *The ability to store and index paper document images is crucial for information preservation and retrieval and much effort has been made in this area of document image analysis. This paper presents a method developed for the autonomous clustering and classification of a set of given document images, allowing for a large set of document images to be indexed with minimal human supervision. This method implements a parallel version of an unsupervised, sequential, two-threshold clustering scheme using a standard error distance metric. The classification is based on structural features extracted from each document image. The corpus of images is 225 medical record documents from the USA Department of Energy.*

**Keywords:** sequential clustering, document image analysis

## 1 Introduction

Much work has been accomplished in the field of indexing paper documents based on text extracted using OCR methods [4][8][10]. These OCR based systems are often only limited to document retrieval. In many applications it is desirable to have a system that contains robust classifying schemes which capture document relations and structure. In order to incorporate this property, a system must be developed that can create a classification system in which the structure and data are permanently embedded within the document feature representation.

The primary challenge in developing this classification system derives from the unconstrained nature of the documents in need of organization. The corpus of medical records used in our research exists in varying structures and layouts, and no relation can be assumed from one record to the next.

The development of a classification system must take into consideration the unconstrained nature of the corpus of images. Unsupervised classification is another requirement of the system. The ultimate goal is to establish a classification of a corpus of millions of medical records. In this effort, an unsupervised clustering technique was developed and initially tested on a set of 225 randomly selected forms from the medical records.

Clustering has applications in fields such as the life sciences, medical sciences and engineering [2]. There are varying types of clustering algorithms, such as agglomerative clustering, K-means, fuzzy [5], hierarchal and sequential [2][9][10]. Each application has an appropriate clustering scheme and the clustering scheme for our application must accommodate the properties existent in the corpus of millions of document images. The first property is the challenge of handling a large number of data sets (document images). A large number of data sets eliminate the feasibility of a prior knowledge such as the number of clusters (which is necessary for many clustering schemes). The large number of data sets makes a sequential execution more feasible whereas the non-sequential clustering schemes rely on all data sets being stored in memory. The medical record images are not necessarily 'clean' images so robust methods were developed to best accommodate this unconstrained property.

In our application, the classification system was developed using a parallel version of an unsupervised, sequential clustering algorithm implementing a two threshold scheme [9]. The sequential scheme does not require that the information of each data set be in memory concurrently and does not require a prior knowledge of the data set. The two-threshold scheme provides some remedy to the challenge of robustness. The clustering scheme relies on a calculated metric representing the distance

between two images. This distance metric is a standard error metric [9]. Herein we present the algorithm developed for the classification of document images. A brief introduction will be provided to the creation of the feature vector which is composed of the invariant structural elements of the document image. Next will follow a description of the string matching distance metric, and a description of the two-threshold sequential clustering scheme. Finally, results of the systems performance will be provided. Then a conclusion will follow.

## 2 Feature Extraction

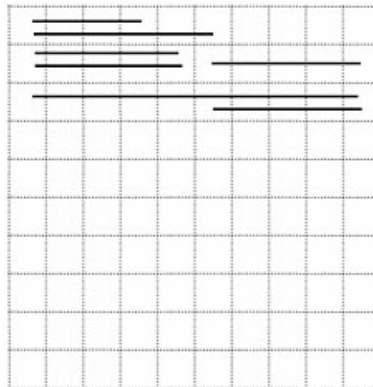
The first step of the classification algorithm is to extract features from the document images. Observation of an arbitrary set of medical records led to the selection of five primary features: form logo, form identification, structural lines (later separated into vertical and horizontal lines), checkboxes and typewritten words. The details of these features have been described in [1]. The focus of this section is to describe how these features are utilized in clustering.

The five primary features are separated into two groups: META features and structural features. The META features are the logo and form identification whereas the structural features include the lines, checkboxes and word locations. The structural features are encoded into a string of 400 elements as seen in (fig 1).

Feature Name	Vector Element(s)
Horizontal Line Crossings	(0...99)
Verical Line Crossings	(100...199)
Checkboxes (center)	(200...299)
Typedwords (center)	(300...399)

Figure 1: The feature vector used in cluster classification

For the positional features (lines, checkboxes and typewritten words), the major body portion of the image is partitioned into a 10 x 10 grid representing 100 elements within the vector. With an image of  $width = w$  and  $height = h$ , each grid element is of size  $\frac{w}{10} \times \frac{h}{10}$ . Each element within the string of 100 elements represents the count for each feature that appeared within that grid element. Typewritten words and checkboxes are represented by a rectangle data structure and their location within the grid is determined by where the center of the rectangles occurs.



(a) line crossing grid

22221100002222211111111112222200000.

(b) result string for grid in (a)

Figure 2: Line crossing grid

For lines (horizontal and vertical), the elements are counted based on crossings. Figure 2 is a 10 x 10 grid with sample lines drawn to simulate actual lines in a document image. A line is traced from its beginning to end and each time the line passes through a new grid element, that element is incremented. The first four grid elements contain the value 2, because there are two lines that have segments passing through those grid elements. Then, only one of these lines continues after the fourth grid to give the elements (5,0) and (6,0) a value of 1. A grid element with no line crossings is given a value of 0.

## 3 Distance Metric

Nearly every clustering algorithm relies on a metric for quantifying the relationship between clusters or cluster representatives. In general, there are two types of metrics: dissimilarity measure (DM) and similarity measure (SM) [9]. The DM is defined as a function  $d$ .

$$\exists d_0 \in \mathcal{R} : -\infty < d_0 \leq d(x, y) < +\infty, \forall x, y \in X \quad (1)$$

where  $x$  and  $y$  are vectors being compared. For our research,  $x$  and  $y$  are the feature vectors of size 400 containing the structural features of a document image described in section 2.

This function contains a lower bound defined at  $d_0 = 0$ . This lower bound is important because when two vectors are exactly the same, the DM should be equal to zero. The higher  $d$  becomes,

the more difference that exist amongst the elements being compared.

The most basic distance measure is the manhattan distance.

$$d(x, y) = \sum_{i=1}^N |x_i - y_i|$$

This distance measure can be extended to include weights, whereby the weights are based on the probability distribution of the consistency of each feature

$$d(x, y) = \sum_{i=1}^N w_i |x_i - y_i|$$

This difficulties associated with this measure is assigning weights to each of the elements of the feature vector. The other basic distance measure is the Euclidean distance

$$d(x, y) = \sqrt{\sum_{i=1}^N (x_i - y_i)^2}$$

The Manhattan and Euclidean distance measures provide for satisfactory distance measures. The challenge arises when working in an unconstrained such as the document images in our research. Algorithms have been developed to reduce noise and create as clean image as possible, but in the end, an image is so degraded that small pieces of information are missing. For example, a line element,  $x_i$ , has 34 lines passing through it and  $y_i$  has 40 lines passing through it. The difference is 6 but is only a 15% error. In another case,  $x_i$  may be 1 and  $y_i$  may be 7, and again, the difference is 6 but the error is 85%. This error is not understood by the Manhattan or Euclidean norms. So, a measure is needed that can incorporate robustness and represent error.

The DM we chose for the clustering scheme is a square error function [9]. This function,  $d(x, y)$ , is

$$d(x, y) = \sqrt{\frac{1}{N} \sum_{i=1}^N \left( \frac{x_i - y_i}{x_i + y_i} \right)^2} \quad (2)$$

$d(x, y)$  returns a result of zero if the vectors are equal.  $d(x, y)$  will reflect the difference along with the percent error of the two vectors by becoming larger as the two vectors grow in difference.

This metric provides a reliable characteristic representing the difference that exists between images, and in our implementation, achieves satisfactory results.

## 4 Unsupervised Classification

The system we developed has the ultimate goal of performing a classification of millions of document images. So, the classification algorithm we choose must be scalable. We also wanted an algorithm that could be parallelized with relative ease. It is not practical that one machine running a sequential algorithm would classify millions of document images. If it takes a machine an average of 10 seconds to classify one image, then it would take 350 days for that machine to classify 3 million images.

So, we developed a classification algorithm which is easily parallelized. The algorithm is based off of a Two-Threshold sequential algorithm [9].

### 4.1 Two-Threshold Sequential Algorithm

The algorithm we developed for classification is derived from a two-threshold algorithm which is a form of a basic sequential clustering algorithm. A sequential algorithm is chosen due to the dynamic nature of the potential application and even more so, for scalability. The ultimate goal is to have the clustering algorithm classify millions of document images. The scale of this problem makes it impossible to store all feature vectors in memory thus creating the necessity of sequential classification. The following is the basic sequential algorithm:

```

BASIC SEQUENTIAL ALGORITHM(imageSet)
1   $n \leftarrow \text{numberOfClusters} \leftarrow 1$ 
2   $x \leftarrow \text{imageSet}_0$ 
3   $\text{cluster}_n \leftarrow x$ 
4  for  $i \leftarrow 2$  to  $N$ 
5      do  $x \leftarrow \text{imageSet}_i$ 
6           $C_{min} \leftarrow \min_{1 \leq j \leq n} d(x, \text{cluster}_j)$ 
7          if  $d(x_i, C_{min}) > \Theta$ 
8              then  $n = n + 1$ 
9                   $\text{cluster}_n \leftarrow x$ 
10
11         else  $\text{ADD}(\text{cluster}_{min}, x_i)$ 

```

The basic sequential algorithm iteratively processes the feature vector of each input image and makes a determination whether the current image is a member of a cluster or constitutes a new clus-

ter. This decision relies upon the selection of a single threshold value and this value is user-specified. This threshold can be determined through experimentation until an acceptable bound is found.

An additional essential selection that must be made by the user is the choice of the distance measure,  $d(x, C)$ . The choice of the distance threshold is directly related to the choice in  $d(x, C)$  wherein different  $d(x, C)$  mean different ranges of difference values.

Along with choosing an appropriate  $d$ , a decision must be made as to a representation for the cluster in the distance calculation. One such method could be a minimum selection over all items within the cluster.

$$\arg \min_{y \in C} d(x, y) \quad (3)$$

The maximum and average are additional alternative calculations. These methods require the use of a distance metric be calculated for each item in the cluster. This becomes computationally expensive for clusters that are large and domains in which there are numerous clusters. So, an alternative is to determine a cluster representative. Several techniques have been developed for finding a cluster representative [2]. This technique requires only one distance calculation to determine a particular sample image's relation to the entire cluster. This advantage greatly reduces the computational time needed for this step. The system we developed uses a representative,  $r$ , of a cluster,  $C$ , where

$$r_i = \frac{1}{N} \sum_{i=1}^N x_{ji}, \quad \forall j \in C \quad (4)$$

where  $r$  is the same size as the feature vectors of  $C$ .

Besides the user-specified threshold values, this algorithm suffers another major pitfall. The order in which the images are presented is very important. A different ordering of the images may produce a different clustering, either in the number of clusters or the clusters themselves. These two properties of this algorithm are not acceptable in the classification scheme necessary for this application.

Thus, a two-threshold algorithm is presented [9] that suppress the two major pitfalls of a simple clustering algorithm. This new algorithm, presented in figure 5, employs a new threshold. The first threshold  $\Theta_1$ , acts a lower bound whereas the second threshold  $\Theta_2$ , acts as an upper bound. Much

like the first algorithm, if  $d(x, C)$  less than  $\Theta_1$  for some  $x$ , then  $x$  will be assigned to  $C$ , whereas if  $d(x, C)$  is greater than  $\Theta_2$ ,  $x$  will be assigned to a newly created cluster. The primary difference between the basic sequential clustering algorithm and the two-threshold algorithm lies the in region between  $\Theta_1$  and  $\Theta_2$ . If  $\Theta_1 \leq d(x, C) \leq \Theta_2$ , then  $x$  is assigned at a later stage. This region is considered a "gray" region in which the relationship between a vector and a cluster is undetermined and will be calculated in a future iteration.

Like the basic sequential algorithm (figure 4), values for the threshold must be user defined, but the additional threshold decreases the impact of choosing a poor value for the threshold. In most cases, experimentation and observation of results is sufficient for deciding values for  $\Theta_1$  and  $\Theta_2$ .

Another appealing property of the two-threshold algorithm is the decreased dependence on order. When a vector occurs within, then that vector's assignment is postponed for a future iteration of the while loop. This assignment does not occur until enough information is available.

## 4.2 Parallel Two-Threshold Classification

We developed an algorithm that could be easily parallelized. This algorithm incorporates the strengths of the Two-Threshold Sequential Algorithm and mechanisms to eliminate the weaknesses. A weakness with the Two-Threshold Sequential Algorithm is the dependency in the order in which the images are presented. This algorithm incorporates a sequential update phase to classify images that were left unclassified by a client process. The other weakness is the lack of scalability of the Two-Threshold Sequential Algorithm. The algorithm works on a closed set of  $N$  images whereas the parallel algorithm we developed is designed to run forever using a dispenser mechanism to provide for any number of clients.

The algorithm is shown in figure 3. The state,  $S$ , of the entire system is a  $X$ -tuple defined as

$$S(D, C, I, E, d, cl, u)$$

where  $I$  is a nonempty set of document image and  $D$  is the storage of document images where  $D = I \cup \{\emptyset\}$ .  $E$  represents a cluster entry in the database  $C$  where  $C = E \cup \{\emptyset\}$ . The process  $d$  is the dispenser and is implemented as a server and its interaction is with  $D$  and the unconstrained set of client processes  $cl \neq \{\emptyset\}$ .  $d$ 's action is defined as

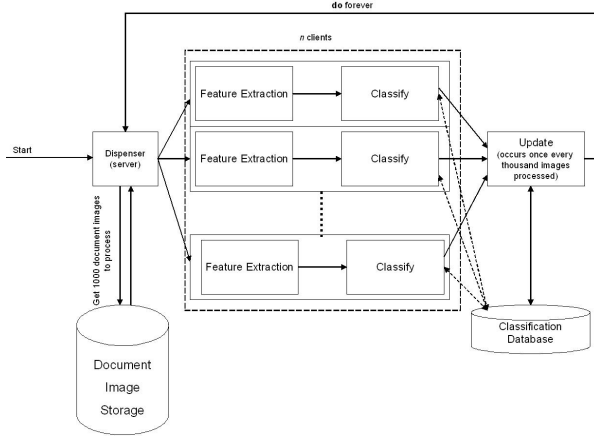


Figure 3: The classification process.

$$d \times D \longrightarrow \{cl\}$$

where  $d$  retrieves a set of  $N$  document images from  $D$  and then distributes amongst the full set,  $\{cl\}$ , of clients. This distribution is based on a request/send protocol, where a client,  $cl_i$  ready to process an image, sends a request to  $d$  and  $d$  sends an image to be processed to  $cl_i$ . The dispenser,  $d$ , does this until all of its retrieved  $N$  images have been sent to be processed by a client.

The client process,  $cl_i$ , has three interactions. One is the input,  $I$ , it receives from the dispenser,  $d$ , and the next two are outputs. One is to write an entry,  $E$ , to the database,  $C$ , and the second is a result message to the update process  $u$ .

Each client executes a simple version of the Two-Threshold Sequential classification algorithm to classify an image. The algorithm for this process is

```

CLIENT TWO-THRESHOLD ALGORITHM( $x$ )
1  $C_{min} \leftarrow \min_{1 \leq j \leq \text{numberOfClusters}} d(x, \text{cluster}_j)$ 
2 if  $d(x_i, C_{min}) < \Theta_L$ 
3   then  $C_{min} \leftarrow C_{min} \cup \{x\}$ 
4      $\text{clusterID} = C_{min}.id$ 
5     return  $\text{clusterID}$ 
6
7 else if  $d(x_i, C_{min}) > \Theta_U$ 
8   then  $\text{numberOfClusters} ++$ 
9      $\text{cluster}_{\text{numberOfClusters}} \leftarrow x$ 
10     $\text{cluster} = \text{numberOfClusters}$ 
11    return  $\text{clusterID}$ 
12
13 else add unclassified
14    return unclassifiedID

```

The Two-Threshold Algorithm for the clients saves the task of assigning those images that appear in the "gray" area to the update process,  $u$ . Each processor client,  $cl$ , sends a message to the update process,  $u$ , upon completion. Once the update process has received  $N$  messages, the update protocol is initiated. The update protocol is

```

CLASSIFICATION UPDATE PROTOCOL()
1  $U \leftarrow$  all unclassified documents
2  $N \leftarrow$  size of U
3 for  $i \leftarrow 1$  to  $N$ 
4   do  $C_{min} \leftarrow \min_{1 \leq j \leq |C|} d(x_i, \text{cluster}_j)$ 
5     if  $d(x_i, C_{min}) < \frac{\Theta_L + \Theta_U}{2}$ 
6       then  $C_{min} \leftarrow C_{min} \cup \{x\}$ 
7          $\text{clusterID} = C_{min}.id$ 
8     else  $\text{numberOfClusters} ++$ 
9        $\text{cluster}_{\text{numberOfClusters}} \leftarrow x$ 
10       $\text{cluster} = \text{numberOfClusters}$ 

```

where  $|C|$  is the number of clusters. The update protocol uses a bound of  $\frac{\Theta_L + \Theta_U}{2}$  to loosen the constraint on membership to a cluster. This helps the robustness of the system. Some documents may have qualities that prevent the complete extraction of features, such as missing information and loss of structural integrity due to image degradation for example. Once the update process finishes, then a message is sent to the dispenser and the system begins again with the dispenser getting another  $N$  images.

## 5 Implementation and Results

The clustering system developed at the Digital Image Processing Laboratory at UNLV, implements the parallel version of the two-threshold sequential clustering scheme. For  $d(x, C)$ , the clustering scheme implements a standard error sum discussed in section 3. The clustering algorithm was initially tested on a corpus of 225 document images. These images are selected from the medical records of the Department Of Energy and were chosen as a small unconstrained subset of the complete corpus of documents.

In order to test the two-threshold algorithm, the feature vectors for the 225 document images were first extracted. Then, observations were made of the values for  $d(x, C)$  between the individual vectors. The observation is an important step in implementing the two-threshold algorithm because the

user must select values for  $\Theta_1$  and  $\Theta_2$ . Choosing appropriate values for  $\Theta_1$  and  $\Theta_2$  may require selecting different variations for each threshold and then performing a clustering operation on a data set. Through observation and testing we found an acceptable range for  $\Theta_1$  and  $\Theta_2$ . Presented in our results are the true clusters computed manually, the result of the Sequential Two-Threshold Clustering algorithm (STTC), and the results of the parallel version of the Two-Threshold clustering algorithm (PTTC).

Table 1: Distribution of Clusters

Cluster (members)	Size	Number of Clusters of this size
1	102	
2	12	
4	1	
5	3	
8	1	
12	1	
18	1	
19	1	
20	1	

The test set of 225 images were manually organized into their correct clusters. Table I shows the sizes of the clusters and how many clusters are of each size. There are a total of 123 clusters.

The error in the STTC is 42.3%. Examining the data results, the primary cause of this difference is high variability amongst images of the same type. Document images of the same type can contain a large distance due to poor scanning and poor image quality. Several methods are being developed to decrease the effect that poor scanning and image quality will have on the classification system. One method targets the design of the distance metric (section 3). The primary feature of the proposed metric would be one that increases the contribution of locations that are the same and make it more difficult for differences to contribute to the overall difference. This metric will allow for small amounts of differences to contribute little to the distance whereas differences will have to be large to have significant impact on the metric.

The error from STTC to PTTC is 31.5%. The primary contribution to this error is primarily due to the difference in which "gray" area images are handled. STTC uses a multi-pass algorithm whereas the PTTC processes all unclassified images separate from the classification process and PTTC does unclassified process in one pass. In addition,

Table 2: Cluster counts for the two algorithms

Cluster (members)	Size	Real Value	STTC	PTTC
1	102	84	103	
2	12	19	18	
3	0	10	6	
4	1	7	1	
5	3	1	1	
6	0	3	0	
7	0	2	0	
8	1	1	0	
9	0	0	1	
10	0	0	0	
11	0	0	0	
12	1	0	0	
18	1	0	2	
19	1	0	0	
20	1	0	0	
21	0	0	0	
22	0	0	0	
23	0	0	1	
Total	123	127	133	

the test on the set of 225 images has no dependence on order. This method is being tested on sets of 10,000 images and more and in these cases, since the update method is called multiple times, there is a dependence on order. Future work in this process will include CLUSTER\_MERGE and CLUSTER\_SEPARATE operations. These methods will eliminate the dependence on order because many clusters will be formed that are closely related and need to be merged, and cluster will be formed with images that do not belong and need to be separated.

Future work also includes the incorporation of more sophisticated classification methods. This area is an important research topic since developing a clustering algorithm which has no knowledge of the number of clusters and no knowledge of how many elements are being clustered could have vast applications.

## 6 Conclusion

The challenge of classifying for indexing large sets of document images is comparable to that of challenges in other fields, such as market analysts who wish to organize large sets of shoppers into like groups. Astronomers group large sets of stars together based on their properties. Each of these applications and many more rely on statistical clus-

tering to provide an unsupervised method for their unique groupings. So, in order to achieve unsupervised classification of document images, a clustering system was developed.

This paper presented a system developed to classify a small corpus of document images. This system differs from the designs of other document indexing systems in that our system uses the invariant structures of the document as the features for classification.

There are five primary invariant features that are extracted from each document image (document logo, document identification, structural lines, checkboxes and typewritten words). These features are then encoded into a vector. This vector is used for calculating the distances necessary for the clustering scheme.

Various choices can be made in choosing a distance calculator. The two general types of measures are a dissimilarity measure and a similarity measure. We chose to develop a dissimilarity measure using a standard error calculation. The distance measure we implemented sufficiently reflected the relationship between two vectors. The implementation of a dissimilarity measure is of future interest and research is being devoted to developing a dissimilarity measure that can accurately distinguish documents that are closely related.

The chosen clustering scheme is a parallel version of a sequential algorithm that employs two-thresholds. This scheme performed well for using the standard error calculation. The clustering algorithm used a minimum matching function (discussed in section 4) to determine the distance between a vector and a cluster. This is an area of improvement and methods such as a Principal Component Cluster representative are being researched and developed.

We are developing methods for the update that will include cluster merging and separation operations along with methods that determine the methods that needs these operations. The near future will include a full scale run on over 3 million document images.

## References

[1] Jae Adams, E.A. Yfantis, D. Curtis and T. Pack. Feature Extraction Methods for Form Recognition Applications. *WSEAS trans. on Information Science and Applications*, Issue 3, Volume 3 March 2006 Pages 666-671.

- [2] M.R. Anderberg. Cluster Analysis for Applications. *Academic Press*, 1973.
- [3] Eugen Barbu, Pierre Héroux, Sébastien Adam, and Eric Trupin. Clustering Document Images Using Graph Summaries. *Lecture Notes in Computer Science*, Volume 3587 Jul 2005 Pages 194-202.
- [4] Thomas Bayer, Ulrich Bohnacker and Ingrid Renz. Information Extraction From Paper Documents. *Handbook of Character Recognition and Document Image Analysis*, 1997 Pages 653-677.
- [5] I. Gath, and A.B Geva. Unsupervised Optimal Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol 11 No. 7 Jul 1989 Pages 773-781.
- [6] Thien M. Ha, and H. Bunke. Image Processing Methods for Document Image Analysis. *Handbook of Character Recognition and Document Image Analysis*, 1997 Pages 1-47.
- [7] J. Hylton. Identifying and Merging related Bibliographic Records. *MIT LCS Masters Thesis*, 1996.
- [8] Kazem Taghva, Julie Borsack and Allend Condit. Information Retrieval and OCR. *Handbook of Character Recognition and Document Image Analysis*, 1997 Pages 755-777.
- [9] S. Theodoridis, K. Koutroumbas Pattern Recognition 2nd Edition. *Academic Press*, 2003 Pages 397-641.
- [10] Yuqiang Guan Large-Scale Clustering: Algorithms and Applications PhD Dissertation University of Texas at Austin 2006