Handwritten Musical Document Retrieval using Music-Score Spotting

ABSTRACT

In this paper, we present a novel approach for retrieval of handwritten musical documents using a query sequence/word of musical scores. In our algorithm, the musical score-words are described as sequences of symbols generated from a universal codebook vocabulary of musical scores. Staff lines are removed first from musical documents using structural analysis of staff lines and their connection with symbols. Symbol codebook vocabulary is created in offline. Next, using this symbol codebook the music symbol information in the document images is encoded. Given a query sequence of musical symbols in a musical score-line, the symbols in the query are searched. Finally, a sub-string matching algorithm is applied to find query words. For codebook, two different feature extraction methods namely: Zernike Moments and Gradient features are tested and two unsupervised classifiers using SOM and K-Mean are evaluated. The performance is measured and compared on a collection of handwritten musical documents.

Keywords
Musical Document Retrieval, Staff Removal, Symbol Classification, Approximate String Matching.

1. INTRODUCTION

Graphics Recognition (GR) has become popular in applications such as Optical Music Recognition (OMR), Engineering Drawing, Maps etc. The main goal is to interpret the graphical documents by recognizing graphical parts and symbols within it. Later, the document contents can be used for efficient indexing according to the interest of the user. With the rapid progress of research in document image analysis and understanding many applications are coming up to manage the paper documents in electronic form to facilitate indexing, viewing, extracting the intended portions etc.

Recently, many works have been done in the analysis of handwritten music scores in context of OMR. The focuses of the research are mainly recognition of handwritten music scores, and the identification of the writer of an anonymous music score [6]. Browsing musical document collection by content information will undoubtedly enhance the user interaction for searching their particular interests. Musical scores could be used as key for searching and indexing of handwritten musical documents. The identification of musical scores in musical document is not easy. It is due to complexity of handwriting, symbol touching with staff lines, etc. The segmentation and recognition of old handwritten music scores is extremely difficult, not only because of the recognition of hand-drawn symbols, but also because of paper aging and degradation. A sample of handwritten musical document is shown in Fig. 1. Due to the difficulties in automatic recognition of hand-drawn music symbols, only the staff removal, writer identification and graphical primitive analysis have been performed. The next steps, mainly recognition, indexing, are still not developed. To the best of our knowledge there has not been any research work on content analysis in musical documents for indexing purpose.

For staff removal algorithms, many research works have been done, since a good detection and removal of the staff lines will allow the correct isolation and segmentation of the musical symbols, and consequently, will ease the correct recognition and classification of the music symbols [3, 4]. Roach and Tatem [5] used a labeling scheme based on the angle information and pixel adjacency to identify these staffline pixels. This approach extracts a number of "horizontal line pixels", some of which belong to music symbols. To avoid the removal of symbol pixels on the stafflines, some horizontal line pixels are

iteratively relabeled as non-horizontal pixels, depending on the labels of their neighboring pixels. Eventually all remaining horizontal pixels are removed. Dalitz et al. [4] have presented a quantitative comparison of different algorithms for the removal of staff-lines from music images. It contains a survey of previously proposed algorithms and suggests a new skeletonization based approach. Concerning the writer identification, musicologists do not only perform a musicological analysis of the composition (melody, harmony, rhythm, etc), but also analyzes the handwriting style of the manuscript. In this sense, writer identification can be performed by analyzing the shape of the hand-drawn music symbols (e.g. music notes, clefs, accidentals, rests, etc) because it has been shown that the author's handwriting style that characterizes a piece of text is also present in a graphic document [14].

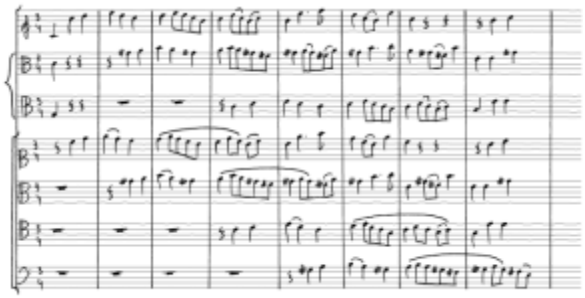


Figure 1: An example of handwritten musical document and its music score symbols with staff-lines are shown.

In this paper we propose a novel approach for retrieval of musical documents based on musical score, which can help in searching or indexing in handwritten or printed historical documents. In our algorithm (See Fig. 2), the musical score-words are described as sequences of symbols generated from a universal codebook vocabulary of musical scores. Staff lines are removed first from musical documents using structural analysis of staff lines and their connection with symbols. Symbol codebook vocabulary is created in offline by a feature extraction and unsupervised classification method. Next, using this symbol codebook the music symbol information in the document images is encoded. Given a query sequence of musical symbols in a musical score-line, the symbols in the query are searched. Finally, an approximate string matching based sub-string matching algorithm is applied to find query words. The best matching classes are selected from system to create a cost matrix to perform the matching. Then the cost matrix is used to find the best match of query sequence in the text sequence.

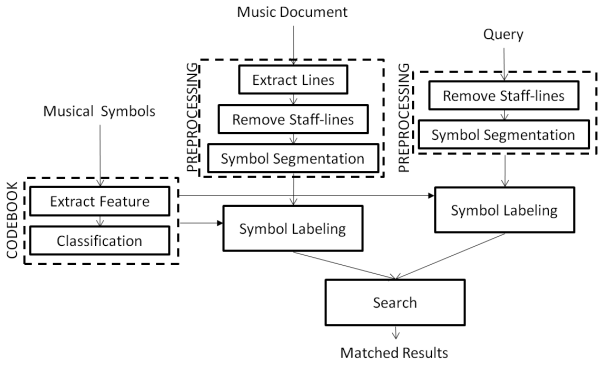


Figure 2: Flowchart of Music Score Retrieval approach in our system.

The main contribution of this paper is to use of dynamic codebook of handwritten music symbols and to generate hypothesis of the music score word location based on the spatial arrangement of these symbols. This approach is efficient to detect music score word in noisy, handwritten document environment. The rest of the paper is organized as follows. In Section 2, we explain the preprocessing and symbol extraction procedure. In Section 3, we describe the feature and classifiers used for codebook creation. Section 4 details the indexing and retrieval process of query music scores. The experimental results are presented in Section 5. Finally conclusion is given in Section 6.

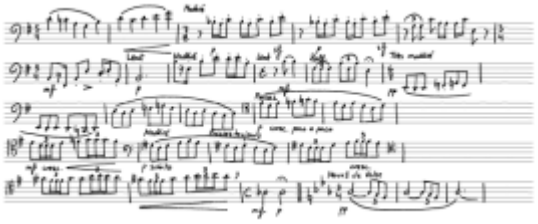
2. PREPROCESSING AND SYMBOL EXTRACTION

2.1 Staff Line Removal

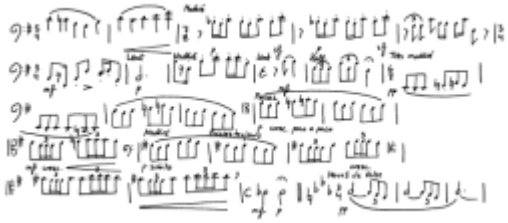
In the literature, almost every paper on OMR deals with the problem and suggests a specific staff removal algorithm [8]. We have used a simple algorithm proposed by [1] for efficient staff line removal from musical documents. The detail of this method is mentioned here in brief.

The music document image is first thinned by skeletonization algorithm. Next, analyzing the thinned image, the thinned line portions are categorized in two groups: (a) straight staff lines and (b) other non-straight or curved staff-lines. Straight lines (part of staff lines) are further divided into horizontal staff lines and non-horizontal straight lines. Next, staff lines are detected based on the characteristics of each group. The staff line detection method can be considered as

passing a ring on a wire (here wire can be considered as staff-line). If there is any obstacle like music score, the obstacle portions is retained or deleted based on some measures. For this purpose, staff line height, staff space height, vertical positional variance of the pixels of thinned lines, etc. are computed and these parameters guide the system to detect the staff line part efficiently. To give an idea of this staff-line segmentation method, we show an image and its staff-line removal result in Fig. 3.



(a)



(b)

Figure 3: Removal of staff-lines in a musical document. (a) A document with staff-lines. (b) Result of staff-line removal

2.2 Symbol Extraction

The musical score lines may contain spurious noise points, irregularities on the boundary of the symbols, etc. Here, background noises are filtered out based on their aspect ratio and pixel density. A major problem of symbol extraction in degraded handwritten documents occurs when the symbols are broken (See Fig.4).

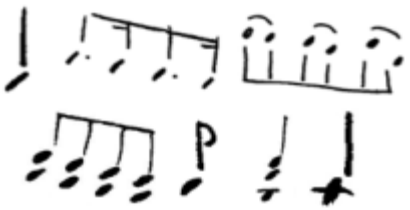
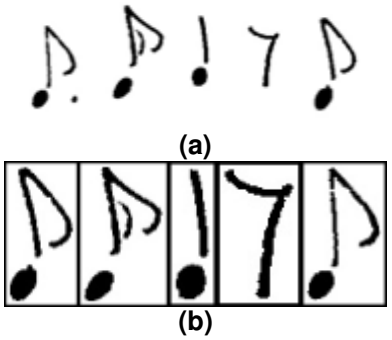


Figure 4: Examples of some broken symbols are shown.

Since, the space between symbols in a musical document is usually much more than that in text document. We have used mathematical morphological operation for joining the broken component. Dilation with 3x 3 structuring element is used to join the broken parts that are very near.

To obtain the individual musical symbols, we apply next a connected component labeling to each musical line image and extract individual components. As a result, the musical scores comprising multiple components will also be segmented into different parts. These components are grouped together by checking their overlapping position horizontally. Two components are grouped together, if one is completely overlapped by the other component or they are overlapped partially by overlapping ratio of Tr. Tr is set to 0.8 according to experiment data. Overlapped components are grouped into a single component. Some musical documents contain vertical straight lines (see Fig.1). These long vertical lines are filtered out next by checking the vertical projection analysis and aspect ratio. The symbol grouping result is shown in Fig. 5.



(a)

(b)

Figure 5: (a) Musical symbols and their segmentation result shown in (b).

3 MUSICAL SCORE CODEBOOK VOCABULARY

A dynamic codebook is created in our system for music score word indexing. Different musical symbols can be represented by a small number of visual components. The representatives are learnt through an unsupervised clustering algorithm of all symbols involved in training.

3.1 Feature Extraction

To take care of scaling effect, in our system two scale invariant features are chosen: Zernike feature and 400-

dimensional feature. These features are used to classify the segmented symbols. The feature extraction is detailed in the following subsections.

3.1.1 Zernike Moment Feature

Zernike moments are based on a set of complex polynomials that form a complete orthogonal set over the interior of the unit circle [9]. They are defined to be the projection of the image function on these orthogonal basis functions. The basis functions $V_{n,m}(x,y)$ are given by:

$$V_{nm}(x,y) = V_{nm}(r,\theta) = R_{nm}(\rho)e^{jm\theta} \quad (1)$$

where n is a non-negative integer, m is non-zero integer subject to the constraints n-|m| is even and $n < |m|$, ρ is the length of the vector from origin to (x, y), θ is the angle between vector ρ and the x-axis in a counter clockwise direction and $R_{n,m}(\rho)$ is the Zernike radial polynomial. The Zernike radial polynomials, $R_{n,m}(\rho)$,are defined as:

$$R_{nm}(\rho) = \sum_{s=0}^{(n-|m|)/2} \frac{(-1)^s (n-s)!}{s!\left[\frac{n+|m|}{2}-s\right]!\left[\frac{n-|m|}{2}-s\right]!} r^{n-2s}$$

Note that, $R_{n,m}(\rho) = R_{n,-m}(\rho)$. The basis functions in equation 1 are orthogonal thus satisfy,

$$\frac{n+1}{\pi}\iint_{x^2+y^2\le 1} V_{nm}(x,y)V^*_{pq}(x,y) = \delta_{np}\delta_{mq}$$

$$\text{where } \delta_{ab} = \begin{cases} 1 \; if & a = b \\ 0 & otherwise \end{cases}$$

The Zernike moment of order n with repetition m for a digital image function $f(x,y)$ is given by

$$Z_{nm} = \frac{n+1}{\pi}\sum_{x^2+y^2\le 1}\sum f(x,y)V^*_{pq}(x,y)$$

where $V^*_{nm}(x,y)$ is the complex conjugate of $V_{nm}(x,y)$.

To compute the Zernike moments of a given image, the image center of mass is taken to be the origin. In our approach, the symbols are normalized into 41*41 before applying Zernike feature computation. The size is considered from the performance of experimental data.

3.1.2 Gradient feature

The gray-scale local-orientation histogram of the component is used for 400 dimensional gradient feature extractions [10]. To obtain gradient features we apply the following steps. At first, size normalization of the input binary image is done. Here we normalize the image into 126 x 126 pixels. The input binary image is then converted into a gray-scale image by applying a 2 x 2 mean filtering 5 times. The gray-scale image is normalized next so that the mean gray scale becomes zero with maximum value 1. Next, the normalized image is segmented into 9 x 9 blocks.

A robust filter is then applied on the image to obtain gradient image. The arc tangent of the gradient (strength of gradient) is quantized into 16 directions (an interval of 22.5) and the strength of the gradient is accumulated with each of the quantized direction. By strength of Gradient (f(x, y)) we mean $f(x,y) = \sqrt{(\Delta u)^2 + (\Delta v)^2}$ and by direction of gradient $(\theta(x,y))$, we mean $(\theta(x,y)) = \tan^{-1}(\Delta u/\Delta v)$, here $\Delta u = g(x+1, y+1) - g(x,y)$, $\Delta v = g(x+1, y) - g(x, y+1)$ and *g(x, y)* is a gray scale value at an (x, y) point. Next, histograms of the values of 16 quantized directions are computed in each of 9 x 9 blocks. Finally, 9 x 9 blocks are down sampled into 5 x 5 by a Gaussian filter. Thus, we get 5 x 5 x 16 = 400 dimensional feature.

3.2 Classifiers

The codebooks are built and tested with 2 unsupervised classifiers: Self Organising Map and K-Means. The classifiers are described in brief below.

3.2.1 Self Organizing Map

The Self-Organizing Map (SOM) [12] is an artificial neural network that performs clustering by means of unsupervised competitive learning. A SOM consists of components called nodes or neurons. Associated with each node is a weight vector of the same dimension as the input data vectors and a position in the map space. The usual arrangement of nodes is a regular spacing in a hexagonal or rectangular grid. The self-organizing map describes a mapping from a higher dimensional input space to a lower dimensional map space. The procedure for placing a vector from data space onto the map is to first find the node with the closest weight vector to the vector taken from data space. Once the closest node is located it is assigned the values from the vector taken from the data space. We show a 4 x 4 SOM map of the music symbols in Fig. 6.

Figure 6: A SOM map (4 x 4) of music symbols.

3.2.2 K-Means

Given a set of observations (x_1, x_2, …, x_n), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets ($k \leq n$) $S = \{S_1, S_2, \ldots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

where μ_i is the mean of points in S_i.

4. RETRIEVAL OF QUERY SEQUENCE

Once the codebooks are created from the segmented training symbols, each musical score lines could be labeled using this codebook vocabulary. Due to the noise and degradation, we do not label the handwritten symbols directly. Instead, we store the feature and use similarity distance measure for detection.

For searching a query word Q from the collection of musical score lines, Q is first segmented into corresponding symbols as explained in Section 2. Next, the music symbols are encoded using the codebook vocabulary. Now, the objective is to find the lines that have similar sequence of symbols. Thus, the matching between query word Q and a target line T is formulated as matching of 2 sequences of symbols.

Approximate string matching (ASM) algorithm [13] has been used in our system for text searching. This method has frequently been used in the literature to refer to a class of pattern matching techniques, by which k errors are allowed between a pattern string Q and a line string T. The length of the strings Q and T may be different. The algorithm finds all substrings of the line T that have at most k errors (character that are not same) with the pattern Q. We use a dynamic cost function (C) for string edit distance computation in ASM algorithm. Computation of C is detailed in the following subsections. Here features of symbols of Q and T are represented as below.

$$Q_i = feature\ of\ i^{th}\ symbol\ of\ Query$$
$$T_j = feature\ of\ j^{th}\ symbol\ of\ Text$$

4.2.1 Cost Matrix Computed in SOM

A SOM of $m \times n$ is trained with features extracted from training symbols as mentioned in Section 3.2.1. The cost function C is computed as follows:

$$C(i,j) = \frac{d_{eu}(V(Q_i), V(T_j))}{d_{max}}$$

where

$$d_{eu}((x_1,x_2),(y_1,y_2)) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2}$$

$$V(S) = \min_{\substack{0 \le x < m \\ 0 \le y < n}} \left(d_{eu}(W(x,y), S)\right)$$

$$W(x,y) = weight\ vector\ of\ SOM\ at\ (x,y)$$

$$d_{max} = d_{eu}(W(0,0), W(m,n))$$

4.2.2 Cost Matrix Computed in K-Means

The symbols are clustered into K classes by K-Mean clustering algorithm. The cluster centres (c_k) are calculated by mean of the elements of the cluster. The distance of feature of a symbol from all the cluster-centres is calculated. A weighted (Gaussian) mean of nearest p vectors is considered as nearest vector. C is calculated as follows:

$$C(i,j) = \frac{d_{eu}(V(Q_i), V(T_j))}{d_{max}}$$

where

$$d_{max} = \max_{\substack{0 \le i < K \\ 0 \le j < K}} \left(d_{eu}(c_i, c_j)\right)$$

$$d_{eu}\left((x_1, x_2, \ldots, x_n), (y_1, y_2, \ldots, y_n)\right) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \cdots + (x_n-y_n)^2}$$

$$V(S) = \sum_{l=0}^{p} w_l(S) * U_l(S)$$

$U_l(S) = l^{\text{th}}\ \text{argmin}_{0 \le i < K}(\text{d}_{\text{eu}}(\text{c}_\text{i}, \text{S}))$,
$w_l(S) =$ weight function computed from Gaussian distribution.
In our experiment, p is chosen as 3 according to experimental results.

5. RESULT AND DISCUSSIONS

5.1 Data collection

We have used the CVCMUSCIMA database [3] to evaluate our system. This dataset contains ground truth of handwritten music score images. It has 1,000 music sheets written by 50 different musicians. Though, the dataset has been especially designed for writer identification and staff removal tasks, in our system these musical scores are used for score-word searching. The groundtruth of score-word level is missing in this dataset. To measure the performance of our system we considered 100 musical document and 15 query samples from this dataset.

5.2 Ground Truth Creation

As the input dataset does not contain ground truth for document retrieval, a semi-automatic tool has been developed to create the ground truth. The groundtruth for 15 queries in 100 documents is created with it. Using this software, the positions of query words are stored as following:

Query# Doc# Line# x1 y1 x2 y2

Where (x1, y1) is co-ordinate of top-left corner of the bounding box of the query and (x2, y2) is the bottom-right corner co-ordinate. Doc# and Line# represent the document name and music line number for that corresponding query.

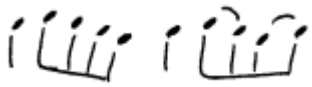
5.3 Performance Metric

To factor in the quality of detection, we consider an outcome correctly detected and complete if the detected region overlaps with more than 75% of the labeled music-score region.

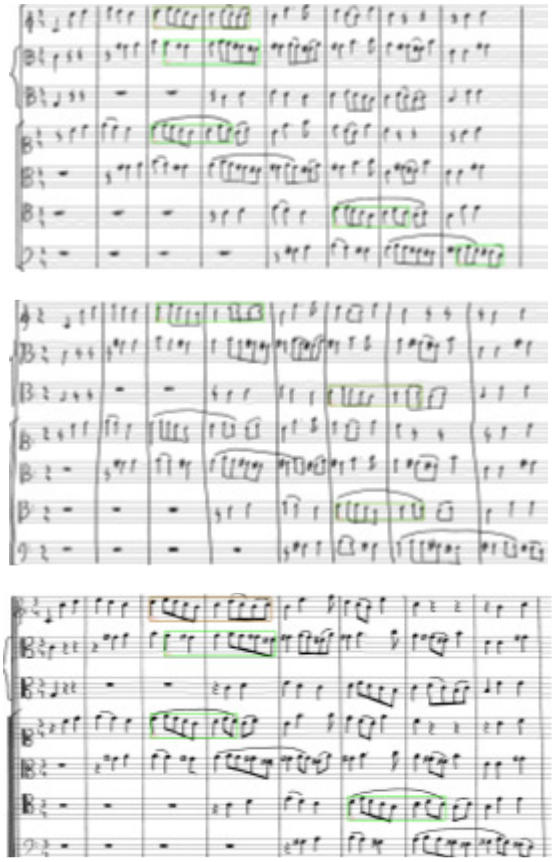
To evaluate the performance of the system with a query word against the collection of musical documents, we use common ratio of precision (P) and recall (R) for evaluation of retrieval of music-word image. The retrieved word images obtained from the system are ranked by the cost obtained in cost matrix. Each result word image is considered as relevant or not depending on the groundtruth of the data. For a given retrieval result, the precision measure (P) is defined as the ratio between the number of relevant retrieved items and the number of retrieved items. The recall (R) is defined as the ratio between number of relevant retrieved items to the total number of relevant items in the collection.

5.4 Qualitative Result

Given a query musical word image, a ranked list of retrieved locations from the database is found. The ranking is done based on accumulated cost in string matching algorithm. To give an idea, we show a query image and its output in Fig.7. Here the outcome of the query is marked with colour. Red and Green indicate the matching result according to rank.



(a) Query



(b) Some of the output results are shown with Red and Green colour

Figure 7: Some retrieval results of a query in musical document image dataset.

5.4 Quantitative Result

In Fig.8, we compare the performance of music word detection system using our approach. Four different

combination results of two features and two classifiers are measured. The precision and recall plot shown in Fig.8 is computed as follows. Based on the cost accumulation for each query musical score image, the retrieved locations are ranked. Finally, we interpolated the results of the precision and recall plot.

From the Fig.8, it is to be noted that the best performance is achieved by the combination result of Zernike feature and K-means clustering. The poor result is due to failure in identification of handwritten musical scores. Since, the documents are degraded and noisy, it reduces the detection accuracy.

5.5 Comparison with DTW Cost Matrix

Dynamic time warping (DTW) is an algorithm for measuring similarity between two sequences which may vary in time or speed. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension.

The profile features [11] are single-valued, i.e. one scalar value is calculated per column in the original image. Here DTW is used on two signals of 4 multidimensional features ($f_k, where\ 1 \leq k \leq 4$): projection profile, upper profile, lower profile and background to foreground transitions.

DTW distance between two signals (features) I_1 and I_2 is calculated using a matrix D.

$$\text{where } D(i,j) = min\begin{pmatrix} D(i,j-1) \\ D(i-1,j) \\ D(i-1,j-1) \end{pmatrix} + d(x_i, y_j)$$

$$d(x_i, y_j) = \sum_{k=1}^{4} (f_k(I_{1,i}) - f_k(I_{2,j}))^2$$

Finally, the matching cost is normalized by the length of the warping path (K). If the length of a feature is l, then the DTW distance between 2 symbols is $D(l,l)$. This cost function $C(i,j)$ is used in ASM algorithm for individual symbol matching.

$$C(i,j) = DTW\_distance(Q_i, T_j)$$

The DTW based cost function is used to find the query words in the database. Fig.8 (yellow line) shows the performance of the DTW based query word retrieval.

We have also tested the system using different number of classes in K-Means clustering algorithm. Fig.9 shows the performance with different number of class numbers. Using K = 15 class, we achieved better performance.

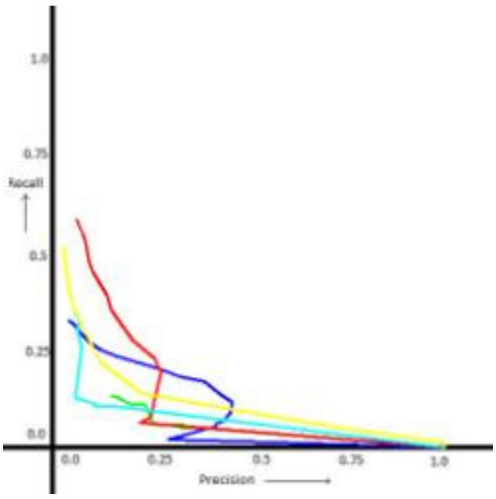


Figure 8: Example of retrieval results of a query in some texts. Here Blue denotes Zernike & K-Means, Green is Zernike & SOM, Red is Gradient Feature & K-Means, Cyan is Gradient Feature & SOM and Yellow is DTW.

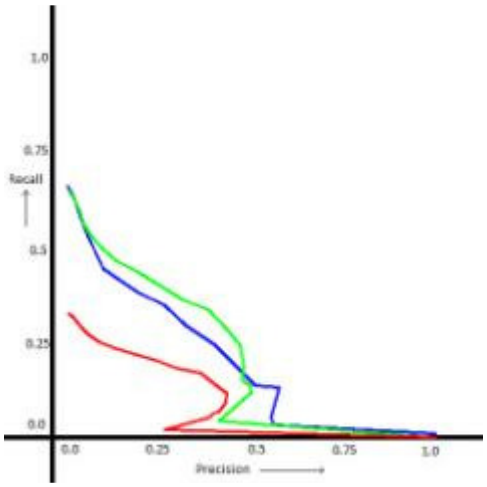


Figure 9: Example of retrieval results Zernike feature and K-Means for different number of classes. Here, Blue denotes 10 classes, Green for 15 classes and Red for 20 classes.

6. CONCLUSION

We have presented here a novel approach of indexing and retrieval for handwritten musical document collection. To extract/identify the musical scores in such unconstrained documents, we have developed

dynamic musical score-codebook based on training symbols. Using visual codebook, the query music-scores is retrieved from each musical score-line from the dataset using a substring matching algorithm. We have tested the performance of our system on codebook vocabulary using two different feature and two unsupervised classifiers. We have also compared our system using Dynamic Time Warping based similarity measures.

The proposed approach works on segmented music lines which are extracted using a simple staff line removal method. The methodology has been made generic and tested in a public dataset. Though the performance results are not satisfactory, we believe it is a step forward for content analysis of handwritten musical documents. To the best of our knowledge, this is the first work of its kind. There are scopes for improvements using this approach by extending the investigation to more accurate segmentation and classification. In future we want to use improved handwritten symbol classifier.

7. REFERENCES

[1] J. R. Chowdhury, U. Pal, "The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification"

[2] S. Li, M. C. Lee, C. M. Pun, "Complex Zernike Moments Features for Shape-Based Image Retrieval", IEEE TSMEC, Part A: Systems and Humans, vol. 39, no. 1, pp. 227-237, 2008.

[3] A. Fornés, A. Dutta, A. Gordo, J. Lladós, "CVC-MUSCIMA: a ground truth of handwritten music score images for writer identification and staff removal", International Journal on Document Analysis and Recognition (preprint), DOI: 10.1007/s10032-011-0168-2.

[4] C. Dalitz, M. Droettboom, B. Pranzas, I. Fujinaga, "A Comparative Study of Staff Removal Algorithms," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 5, pp. 753-766, 2008.

[5] J.W. Roach, J.E. Tatem: Using Domain Knowledge in low-level visual Processing to interpret handwritten Music: an Experiment. Pattern Recognition 21, pp. 33-44, 1988.

[6] A. Fornés, J. Lladós, G. Sánchez, X. Otazu, H. Bunke, "A combination of features for symbol-independent writer identification in old music scores", IJDAR 13(4), pp. 243-259, 2010.

[7] M. V. Stuckelberg, D. Doermann, "On musical score recognition using probabilistic reasoning", Proceedings of the 5th International Conference on Document Analysis and Recognition, pp. 115-119, 1999.

[8] D. Blostein, H.S. Baird: A Critical Survey of Music Image Analysis. In H.S. Baird, H. Bunke, K. Yamamoto (editors): "Structured Document Image Analysis", pp. 405-434, Springer, 1992.

[9] S. Li, M. C. Lee, C. M. Pun, "Complex Zernike Moments Features for Shape-Based Image Retrieval", IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, vol. 39, no. 1, pp. 227-237, 2008.

[10] U. Pal, N. Sharma, T. Wakabayashi, and F. Kimura, Handwritten numeral recognition of six popular Indian scripts," In Proc. International Conference on Document Analysis and Recognition, pp. 749-753, 2007.

[11] T. M. Rath, R. Manmatha, "Features for Word Spotting in Historical Manuscripts,", International Conference on Document Analysis and Recognition, pp.218 -222, 2003

[12] Self Organising Map (SOM) http://en.wikipedia.org/wiki/Som

[13] P. P. Roy, J.Y. Ramel and N. Ragot, "Word Retrieval in Historical Document using Character-Primitives", International Conference on Document Analysis and Recognition, pp. 678-682, 2011.

[14] A. Fornés, J. Lladós: A Symbol-Dependent Writer Identification Approach in Old Handwritten Music Scores. ICFHR, pp. 634-639, 2010.