Handwritten Musical Document Retrieval using Music-Score Spotting

Rakesh Malik
School of Engineering &
Technology, Pondicherry
University, India
rakeshmalik91@gmail.com

Partha Pratim Roy
CVPR Unit,
Indian Statistical Institute,
Kolkata, India
2partharoy@gmail.com

Umapada Pal
CVPR Unit,
Indian Statistical Institute,
Kolkata, India
umapada@isical.ac.in

Fumitaka Kimura
Graduate School of Engg.,
Mie University, Mie,
Japan
kimura@hi.info.mie-u.ac.jp

***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 symbol codebook vocabulary is created in offline. Next, using this symbol codebook the music symbol information in each document image is encoded. Given a query sequence of musical symbols in a musical score-line, the symbols in the query are searched in each of these encoded documents. Finally, a sub-string matching algorithm is applied to find query words. For codebook, two different feature extraction methods namely: Zernike Moments and 400 dimensional gradient features are tested and two unsupervised classifiers using SOM and K-Mean are evaluated. The results are compared with a baseline approach of DTW. The performance is measured on a collection of handwritten musical documents and results are promising.**

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

I. INTRODUCTION

Graphics Recognition (GR) has become popular in applications such as Optical Music Recognition (OMR), Engineering Drawing, Maps etc. The main goal of GR 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. 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. 2(a). Due to the difficulties in the automatic recognition of hand-drawn music symbols, only the staff removal, writer identification and graphical primitive analysis have been reported in the literature. There is a lack of research effort towards the work like recognition, indexing etc. To the best of our knowledge there has not been any research work on content analysis in musical documents for indexing purpose.

Many research works have been done for staff removal algorithms, 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 staff-line 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 staff-lines, 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. The approach proposed by Fornes et al. [3] does 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, 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.

In this paper we propose a novel approach for retrieval of musical documents based on musical score spotting, which can help in searching or indexing handwritten or printed historical documents. In our scheme 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 extracting different features and using 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. Flow diagram of the proposed scheme is shown in Fig.1.

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

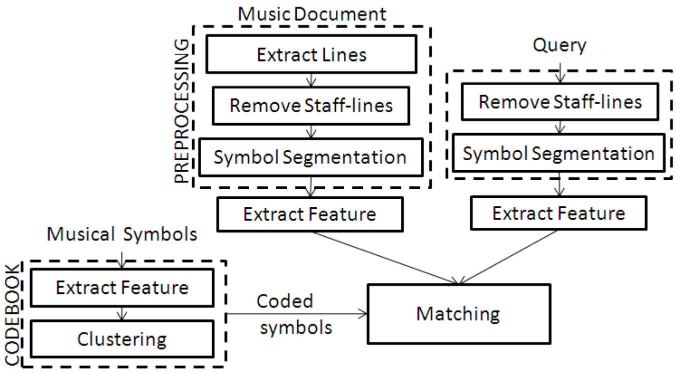


Figure 1: Flowchart of Music Score Retrieval approach.

II. Preprocessing and Symbol Extraction

A. Staff Line Removal

Staff Line removal from musical score is an important and challenging task. Here we have used the algorithm due to Chowdhury and Pal of [1] for efficient staff line removal from musical documents. This algorithm obtained the highest results in staff Line removal competitions of ICDAR-2011 [1]. 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. 2.

B. Symbol Extraction

The musical score lines may contain spurious noise points, irregularities on the boundary of the symbols, etc. These noises are filtered out based on their aspect ratio and pixel density. Major problems of symbol extraction in degraded handwritten documents occur when the symbols are broken (See Fig.3). 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 a proper structuring element is used to join the broken parts that are very near.

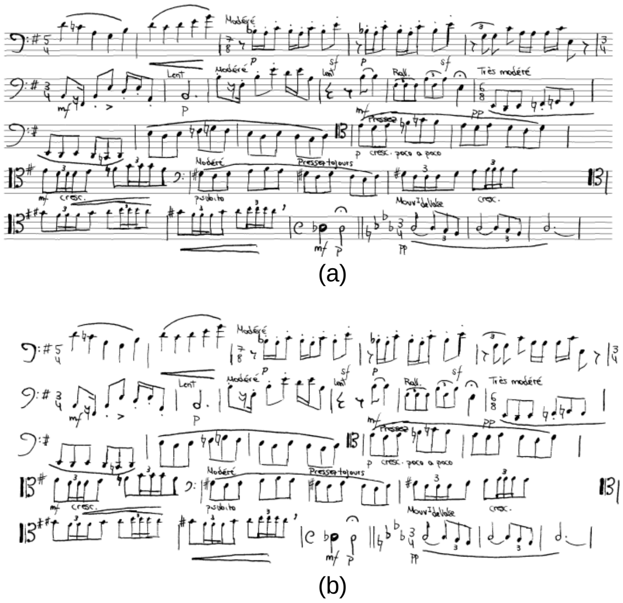

(a)
(b)

Figure 2: Examples of staff-line removal in a musical document. (a) A document with staff-lines. (b) Result of staff-line removal.

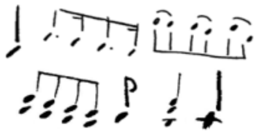


Figure 3: Examples of some broken symbols.

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 vertically. Two components are grouped together, if one is completely overlapped by the other or they are overlapped partially with overlapping ratio greater than *Tr*. *Tr* is set to 0.7 according to experiment data. Overlapped components are grouped into a single component. The symbol grouping result is shown in Fig. 4.

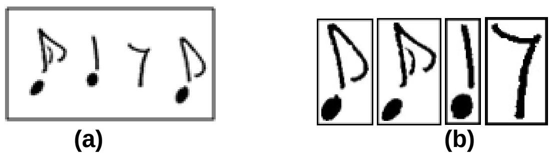

(a)
(b)

Figure 4: (a) Musical symbols and their segmentation result are shown in (b).

III. 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.

A. *Feature Extraction*

To take care of scaling effect, in our system two scale invariant features are chosen: Zernike feature and 400-dimensional gradient feature. These features are used to classify the segmented symbols in our approach. The feature extraction process is detailed in the following subsections.

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:

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

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. To compute the Zernike moments of a given image, the image center of mass is taken to be the origin.

2) Gradient feature: The gray-scale local-orientation histogram of the component is used for gradient feature extractions [10]. The image is normalized into 126x126 size and converted to gray-scale image by applying a set of mean-filtering. Next the resultant gray image is segmented into 9X9 blocks. Roberts filter is applied next to obtain gradient image. The direction of gradient is quantized into 16 directions and the gradient strengths are accumulated in each quantized direction. Histograms of 16 quantized directions are computed in each of 9x9 blocks. Finally, 9x9 blocks are down sampled into 5x5 by a Gaussian filter. Thus, we get 5x5x16 = 400 dimensional feature.

B. *Classifiers*

The codebooks are developed by unsupervised classifiers. In this present work, we have tested 2 unsupervised classifiers: Self Organising Map and K-Means. The classifiers are described in brief below.

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. We show a 4 x 4 SOM map of the music symbols in Fig. 5.

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 ≤ n) S = {S_1, S_2, …, S_k} so as to minimize the within-cluster sum of squares (WCSS):

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

where μ_i is the mean of points in S_i.

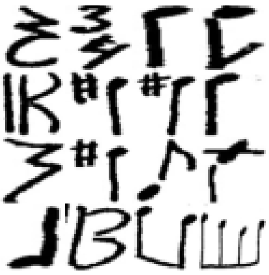


Figure 5: SOM map (4 x 4) of music symbols.

IV. INDEXING AND RETRIEVAL

A. *Indexing of Musical Documents*

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. This is detailed below.

B. *Retrieval of Query Music word*

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 symbols are encoded using the codebook vocabulary. Now, the matching between query word *Q* and a target *T* is formulated as matching of 2 sequences of symbols.

Approximate string matching (ASM) algorithm [13] has been used in our system for searching of sequence of symbols. 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 *T*. The cost function (*C*) for string edit distance is computed from cost matrix of similarity measures. *C* is used in ASM algorithm to find the best matching of query string into the line string. Computation of *C* is explained in details in the following subsections. Here features of symbols of *Q* and *T* are represented as follows:

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

1) Cost Matrix Computed in SOM: A SOM of $m \times n$ is trained with features extracted from training symbols as mentioned in the last Section. A matrix *C* of order $|Q| \times |T|$ (where |.| indicates the length) is computed as follows:

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

where

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

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

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

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

2)Cost Matrix Computed in K-Means :During searching *of* a query word, the features of the components of the query (Q) and target (T) of length m and n respectively are obtained. The system is trained with features of many symbols and clustered into N classes. The cluster centres 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 l vectors is considered as nearest vector. A matrix C of order $|Q| \times |T|$ is calculated as follows:

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

where

$$d_{max} = \max_{\substack{0\le i<|Q|\\0\le j<|T|}} \left(d_{eu}(Q_i,T_j)\right)$$

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

$$V_{mean}(S) = \sum_{u=0}^{l} w_u(S) * V_u(S)$$

$V_u(S) = \mathrm{u^{th}} \underset{0\le i<N}{\mathrm{argmin}}(\mathrm{d_{eu}}(\mathrm{c_u},\mathrm{S}))$,
$w_u(S) = weight$ function, $c_u = centre\ of\ u^{th}\ class.$

V. RESULT AND DISCUSSIONS

A. Dataset

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 musical 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.

B. Ground Truth Creation

As the input dataset taken 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 in a file named Doc# as following:

```
Query# Doc# Line# x1 y1 x2 y2
```

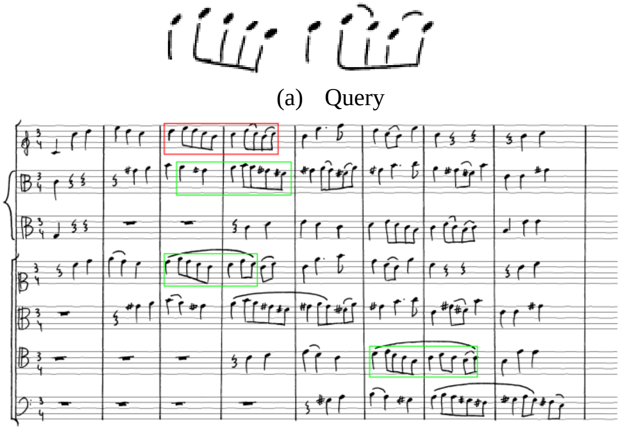
Where (x1, y1) is co-ordinate of top-left corner of the bounding box of the query and (x2, y2) is the bottom-right corner co-ordinate. Doc# and Line# represent the document name and music line number for that corresponding query.

C. Performance Metric

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 the number of relevant retrieved items to the total number of relevant items in the collection.

D. Qualitative Result

Some of retrieval results are shown in the following figures. Here the queries are marked with colours. The Red colour indicates the best match and the Green indicates bad match.



(a) Query

(b) Retrieval result for the shown query (See the PDF file for better visibility)

Figure 6: Some retrieval results of a query in musical document image.

E.Quantitative Result

Fig. 7 shows the precision-recall curve for different feature and classifiers used in our experiment. The best result is obtained using Zernike feature and K-Mean Clustering. The Zernike feature was of dimension 41. From the curve, it shows the Zernike features provides better performance than 400 dimensional gradient features. In Fig. 8, we showed the performance of K-Means clustering with different number of clusters chosen. The Zernike features is used in this test for its superior performance. With K= 20 clusters we obtained best result.

F. Comparison with DTW Cost Matrix

Here, we compare our algorithm with a baseline approach of Dynamic time warping (DTW) which is popular for its word spotting technique. 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. DTW is used on two signals of 4 multidimensional features ($f_k, where\ 1 \leq k \leq 4$): projection profile, upper and lower word profile and foreground transition. 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). This cost function $C(i,j)$ is used in ASM algorithm for individual symbol matching.

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

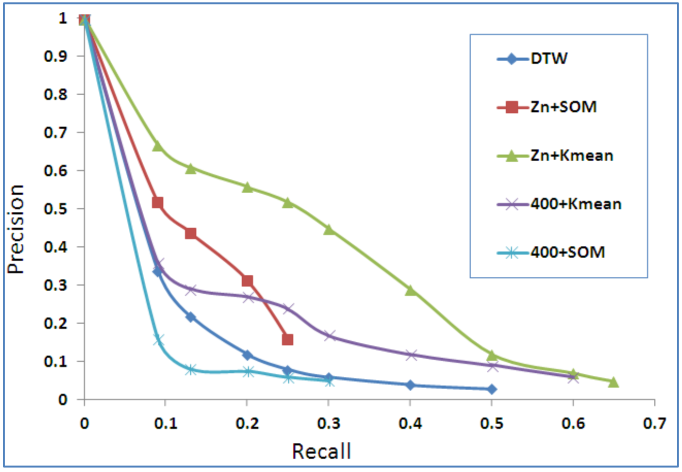


Figure 7: Performance of retrieval results with the combination of features and classifiers. These combined methods are compared against DTW. Zn denoted Zernike feature.

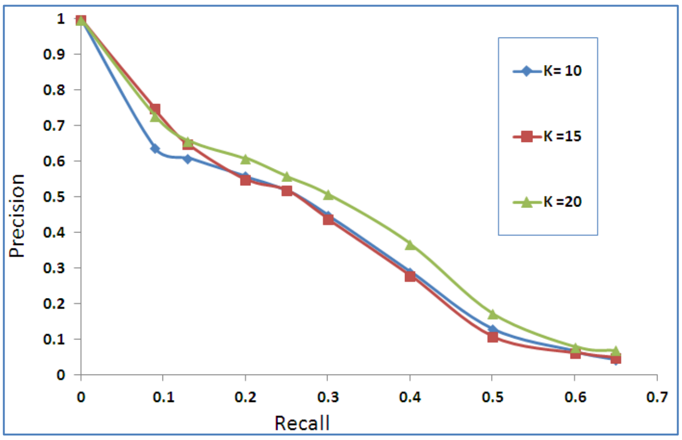


Figure 8: Performance of retrieval results using Zernike feature and K-Means clustering. Different K values are tested in our experiment.

VI. CONCLUSION

We have presented here a novel approach of indexing and retrieval for handwritten musical document collection. To extract/identify the musical scores in unconstrained documents, we have developed dynamic musical score-codebook based on training symbols. Based on the codebook, the music-scores in each line are indexed. The retrieval is performed using a substring matching algorithm using the dynamic codebook. We have evaluated our system using two different features and two unsupervised classifiers. The similarity measures in string matching are performed using Euclidean Distance and Dynamic Time Warping.

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. There are scopes for improvements using this approach by extending the investigation to more accurate segmentation and classification. In future we want to extend this system using improved handwritten symbol classifier.

REFERENCES

[1] Alicia Fornes, Anjan Dutta, Albert Gordo and Josep Llados, "The ICDAR 2011 Music Scores Competition: Staff Removal and Writer Identification", In Proc. ICDAR-2011, pp. 1511-1516, 2011.

[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] Alicia Fornés, Anjan Dutta, Albert Gordo, Josep 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] Christoph Dalitz, Michael Droettboom, Bastian Pranzas, Ichiro Fujinaga, "A Comparative Study of Staff Removal Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 753-766, May 2008, doi:10.1109/TPAMI.2007.

[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] Alicia Fornés, Josep Lladós, Gemma Sánchez, Xavier Otazu, Horst Bunke: A combination of features for symbol-independent writer identification in old music scores. IJDAR 13(4): 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 (ICDAR’99), p. 115 (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: 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 (ICDAR), 2007, 749-753.

[11] Toni M. Rath, R. Manmatha, "Features for Word Spotting in Historical Manuscripts,", International Conference on Document Analysis and Recognition (ICDAR'03) - Volume 1, 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 (ICDAR), pp. 678-682, 2011.