A Covariance-matching-based Model for Musical Symbol Recognition

Luu-Ngoc Do, Hyung-Jeong Yang, Soo-Hyung Kim,
Guee-Sang Lee, Cong Minh Dinh

Abstract

A musical sheet is read by optical music recognition (OMR) systems that automatically recognize and reconstruct the read data to convert them into a machine-readable format such as XML so that the music can be played. This process, however, is very challenging due to the large variety of musical styles, symbol notation, and other distortions. In this paper, we present a model for the recognition of musical symbols through the use of a mobile application, whereby a camera is used to capture the input image; therefore, additional difficulties arise due to variations of the illumination and distortions. For our proposed model, we first generate a line adjacency graph (LAG) to remove the staff lines and to perform primitive detection. After symbol segmentation using the primitive information, we use a covariance-matching method to estimate the similarity between every symbol and pre-defined templates. This method generates the three hypotheses with the highest scores for likelihood measurement. We also add a global consistency (time measurements) to verify the three hypotheses in accordance with the structure of the musical sheets; one of the three hypotheses is chosen through a final decision. The results of the experiment show that our proposed method leads to promising results.

Keywords : musical-symbol recognition|covariance-matching|image processing|pattern recognition|musical rules

I. INTRODUCTION

The optical music recognition (OMR) system is significant for the digitalization of music at minimal cost. It transforms a musical-sheet image into a format such as XML that is readable by a machine so that the file can be converted into a playable electronic format such as a midi file [1]. The following four types of music-score images are those that OMR systems need to handle, and are presented sequentially according to an increasing level of difficulty: printed-scanned images, printed-captured images, handwritten-scanned images, and handwritten-captured images. Since a printed-scanned image is typically very clean with a low noise level, it can be easily recognized by the current state-of-the-art OMR systems [1, 6, 9]. Graph-based models [3, 4, 7] and machine learning-based models [1, 23, 24, 26] are widely used to achieve good performance for this kind of image. A printed-captured image, however, is more challenging because it typically comprises numerous distortions that produce a high level

*Member, Chonnam National University

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (NRF-2017R1A4A1015559) and the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2017-2016-0-00314) supervised by the IITP(Institute for Information & communications Technology Promotion).

Manuscript : 2017 09 04
Revised : 2017. 11. 15

Confirmation of Publication : 2017.11.30
Corresponding Author :Hyung-Jeong Yang, e-mail : hjyang@jnu.ac.kr

of noise in the binarization results. Handwritten-scanned and handwritten-captured musical images present different propositions altogether due to the corresponding variety of indistinct shape that cannot be handled by the use of conventional algorithms.

In this paper, we focus on printed-scanned and printed-captured music-sheet images with the aim of porting the system to a mobile application. To this end, our system provides a reliable performance, even with small changes of the illumination or camera view the final result still be preserved. We therefore propose a covariance-matching-based music score recognition model, whereby the structural information of the musical sheets is embedded

to address the noise and uncertainty. The proposed model uses a covariance descriptor that represents a strong relationship between the spatial information of black pixels to describe the characteristics of each musical symbol. Three hypotheses are extracted to ensure that the true label of the symbols will not be completely removed from the process due to noise, even if it is not the most similar one. The structural information is then used to decide which hypothesis satisfies the musical rule and leads to a final result. We verified the proposed method with printed scanned images and printed-captured images for which we were able to achieve promising results. The contribution of this paper is described as follows:

- We propose a new algorithm for musical-symbol recognition for which a covariance matching is used. This algorithm is supported by a covariance descriptor for the generation of an appropriate result, whereby the musical rules are verified.
- The proposed model is fully automatic and stable. In most cases, even if the conditions of the input images are changed, the output still preserves sound results.
- The parameters dependencies are limited.
- The proposed model can be used for a mobile application.

The remainder of this paper is organized as follows: section 2 briefly describes the state-of-the-art OMR systems as well as their problems and challenges; section 3 presents the proposed OMR system for which covariance matching is used, and includes a brief description of the pre-processing step and the details of the recognition strategy; the results of our experiment are shown in Section 4; and lastly, the conclusion is presented in Section 5.

II. RELATED WORK

The early OMR system was presented by Pruslin [2] and several OMR software packages have appeared in the market over the subsequent years; however, none of these packages have shown a satisfactory performance in terms of precision and robustness. The complexity of OMR tasks is a result of the bi-dimensional structure of musical notation that is formed by staff lines and the existence of several combined symbols that are organized around note heads, and this has hampered progress in the OMR area. Until now, even the most advanced recognition systems such as Smartscore[20] could not identify every musical notations. In addition, the focus of classic OMR systems is directed toward clean printed-scanned music sheets on desktop PCs, and a sound performance is usually obtained only when this type of score is being processed [1].

There are two main approaches for musical symbol recognition in the previous OMR systems: graph-based models and machine learning-based models. The graph-based models tried to construct a graph for each symbol. An analysis of the graph structure was used to classify the musical symbol. The machine learning-based models applied a training process to learn general rules for classification.

In 1992, Carter used a line adjacency graph (LAG) model to extract the symbols [3]. This model generates vertical run-lengths and groups them into sections that can be used for object segmentation. The extracted objects are classified based on their size, the number and organization of their sections. Reed and Parker also used LAG to detect straight lines and

curves of musical symbols while the accidentals, rests and clef are detected by the measurement of perpendicular distance from the object's contour to reference axis [7]. Randriamahefa then proposed a structural method that is based on the construction of graphs for each symbol [4]. To isolate these graphs, a region-growing method and thinning are used. The performance of graph-based models much depends on a pixel-level process which is very sensitive to the change of intensity and image's resolution in the task of recognition. Therefore, recent research on OMR focused on machine learning-based model.

In [23] a Hidden Markov model (HMM) is used to perform segmentation and classification simultaneously. However, this model was applied only for typographic score which was simplified comparing to the modern music score. Homenda and Luckner applied decision tree for a challenged data where the symbols were distorted by noises, printing defects, skew and curvature of scanning [24]. However, this model was designed for recognizing only five classes of musical symbols such as quarter rest, eight rest, sixteenth rest, forte and piano while ignoring most of the important symbols such as black note, white note, flag note and beam note. A fuzzy model was developed by Rossant and Bloch [5, 6] for robust symbol detection and template matching; this method was designed to handle uncertainty, flexibility, and fuzziness at the symbol level. This model generated multiple hypotheses of recognition results to avoid ignoring the true hypothesis due to noises. Musical rules are then used to make decision. Its power,
however, has not been shown for captured images that are essential for the estimation of the performance in a mobile application.

Achieving a comparable performance, Rebelo [1] used a hierarchical decomposition to extract symbols that are then recognized using classifiers such as k-nearest neighbor, neural networks, and support-vector machines (SVM). These classifiers received raw pixels from a 20 x 20 image as input feature. The classification results reported SVM as the best classifier. An elastic deformation method also was presented to handle the problem of symbol variability and imprecision. However, the experimental results showed that this method does not boost the overall accuracy as much as expected. Bellini introduced an object-oriented optical music-recognition system whereby the staff lines are not removed before the segmentation step [8]. A horizontal histogram is generated to vertically slice the symbols into segments, followed by the use of a neural network during the recognition phase. The previously proposed machine learning-based models through the use of training process were reported with very high accuracy. However, the variant of musical symbol shape and uncertainty from captured images are still challenges for these models. Moreover, machine learning-based models also required a normalization of symbols size which can generate the loss of information.

Most OMR systems [1, 5, 6, 9] attempt to recognize symbols by processing pixels or groups of pixels. This kind of process is very sensitive to illumination changes and variations of the shape and scale of the musical symbols. Since the size of the symbols has to be considered through normalization or resizing, the loss of some information can occur. In noisy images, a loss of information critically damages the recognition process; therefore, the use of pixels or groups of pixels for symbol recognition is not optimal for a mobile application. The potential of the covariance descriptor can solve the above issues while it was proved to be invariant with an identical shifting of the intensity value and independence with respect to the size of the symbol's region [10]. In this paper, we employ the covariance to compute the similarity between a symbol and the templates to address such problems.

III. PROPOSED SYSTEM

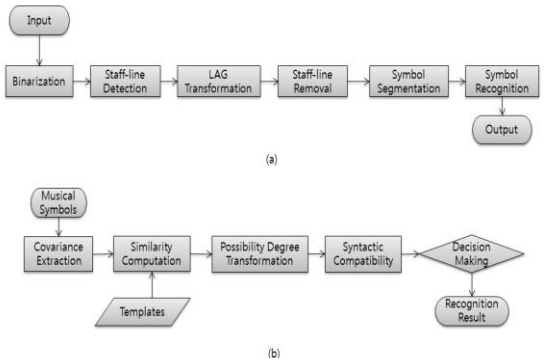

Fig. 1. (a). Proposed OMR system
(b). Symbol Recognition Model

Our system consists of staff-line detection, LAG transformation, staff-line removal, symbol segmentation and symbol classification. In this paper we adopt the local thresholding binarization method which is proposed in [21]. A binarized musical image is used for the detection of the staff-line position. The transformed LAG [3] is generated to group the vertical run-length into sections so that the staff-line removal can be properly performed. After symbol segmentation step, the covariance-matching process extracts the covariance descriptors for each symbol. The distance between two covariance matrices is computed to identify the similarity between these objects and the templates. The similarity is then transformed to a possibility degree and to select three hypotheses. Lastly, the global information is verified with the syntactic rule for each hypothesis so that a decision can be made. The structure of proposed OMR system and the symbol-recognition model is described in Figure 1(a) and Figure 1(b) respectively. In this paper we focus on the symbol recognition step as Figure 1(b).

As a pre-processing step before the extraction of the covariance descriptor, we conduct staff-line detection and removal and symbol segmentation. We follow the boosted stable-path algorithm for staff-line detection, provided in [12] [13] [14], whereby an image is used as a graph and a staff line is considered as a connected path from the left margin to the right margin of the music score. Since the staff lines are essentially the only extensive black objects on the music score, the identified path is the shortest path between the two margins if the entire paths through black pixels are favored.

In the next step, we generate a LAG transformation to remove the staff lines and segment the objects. As described in [22], all of the pixels that belong to the staff lines belong to one section that is separated from sections located inside the musical symbols; accordingly, we can easily remove the staff lines without breaking the musical symbols, as shown in Figure 2(a) and Figure 2(b). The symbol segmentation is based on connected components. Segmentation result should preserve that every symbol should be separated from each other and at most one stem is included. Figure 3 shows some examples of the segmentation results.

After the attainment of the region for each object during the segmentation process, we extract the covariance descriptor. The covariance descriptor was first presented by Tuzel in a problem involving human detection [11] and we adapt this idea to represent the characteristics of any musical symbols in binarized images.

As described in [11], suppose we have a region R for one object in the gray image, the covariance matrix C_R of this object's feature can be computed as follows:

$$C_R = \frac{1}{S}\sum_{k=1}^{S}(f_k - \mu_R)(f_k - \mu_R)^T \qquad (1)$$

where S is the number of pixels in region R, μ_R is the mean of the corresponding features computed from all points of the region R.

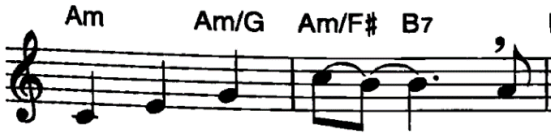

Fig. 2(a). Image before staff-line removal

Fig.2(b). Result of staff-line removal

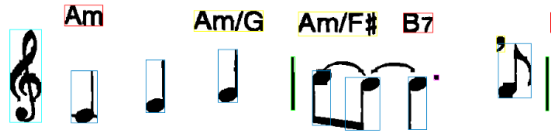

Fig.3. Example of segmentation result

The covariance C_R is a symmetric matrix where the diagonal elements indicate the variances of each feature and other elements indicate their respective correlation. The feature vector f_k can be any kind of mapping from region R such as color, image gradient, and edge orientation. The covariance C_R can be computed from any kind of region, not only the rectangular region, as described in Figure 4.

The covariance matrix of any region has the same size $d \times d$, where d is the size of the

Fig.4. Covariance matrix of region R

feature vector f_k. The covariance matrix therefore represents the characteristics of any region without the need to consider their size or the normalizing feature values. As previously mentioned, the covariance matrix is invariant under varying illumination conditions with an identical shift of the intensity value [11]; moreover, such a property can be maintained even in cases of binary images with small changes of illumination for each camera capturing.

The shapes of musical symbols are usually shown clearer in binary images than gray images. Computing the covariance matrix with only black (or white) pixels in a binary image also saves the cost of process. Therefore, we employ the idea of covariance descriptor to represent musical symbols in binarized images. In a binary image, the extraction of a covariance matrix is slightly different from that of a color or grayscale image in that we only consider black pixels as follows:

$$C_{RB} = \frac{1}{Sb}\sum_{k=1}^{Sb}(f_k - \mu_R)(f_k - \mu_R)^T \quad (2)$$

where Sb is the number of black pixels in region R. Since the shape of an object in the binary image is decided upon by the position of the black pixels, ignoring the white pixels does not generate any loss of information; therefore, the positions of the black pixels or the spatial information are more essential for this process. We choose the first two elements of a feature vector f_k as the coordinates x and y of the black pixels. Even if the values of the elements in the covariance matrix are averaged, the multiplication operator increases them when the size of region R is enlarged. Instead of using the coordinates x and y, we use the ratios x/w and y/h, where w is the width of the symbol's region and h is the height to remove the scaling factor on the diagonal of the covariance matrix; however, the coordinate information alone is insufficient for the characterization of the symbol shape. We also used the distance-map value as one feature of the black pixel since it shows the thickness of a symbol. The distance map of the binary image can be computed by using the distance transform [15]. The values of a distance map indicate the minimum distance from the current black pixel to any other white pixels that follow horizontal and vertical directions. The pixels that belong to the object boundaries are transformed into one, while the pixel of the red cell is transformed into two. The values of the distance map indicate the depth that a black pixel falls into the centroid of a black region, or the distance between the black pixel and the boundary of the object, so that it provides the information of thickness of symbol's shape. For example, the black note and white note have similar shape but the depths of their black regions are different. Many musical symbols have thick shape such as the black note, full and half rest symbols while the other shapes such as white note, whole note and flat are thin. Therefore the information of thickness of symbol shapes is very useful to distinguish musical symbols.

Spatial information of the degree of the vector that connects the black pixel with the origin is considered as the forth feature. The feature vector of our system is as follows:

$$f_k = \left[\frac{x}{w} \quad \frac{y}{h} \quad dst(x,y) \quad \arctan\left(\frac{x}{y}\right)\right] \quad (3)$$

where $dst(x,y)$ is the value of the distance map at (x, y); therefore, the covariance matrix C_{RB} is a 4 x 4 symmetric matrix.

The extracted covariance descriptor can be used to find the best-matched pre-defined templates of a considering symbol. First, we compute the distance between the covariance matrix of the considering symbol and the covariance matrix of the template symbol, whereby the dissimilarity between them is identified by this distance. Next, the distance value is used to generate the possibility degree to indicate the extent of the similarity between the template and the considering symbol. The three highest-possibility templates are chosen as three hypotheses, and lastly the hypothesis that satisfies the syntactic rule with the highest score is chosen as the best matching result. After obtaining the covariance matrices of the

templates and the testing symbols, we compute the distance or dissimilarity between them to find the best match for the classification. A variety of similarity measures are shown in [16, 17, 18]; however, all of them cannot be the optimal way for dealing with the covariance matrices since the covariance matrix does not lie on the Euclidean space [11]. We therefore use the metric proposed by Forstner and Moonen [19] to compute the distance between two covariance matrices:

$$d(C_i, C_j) = \sqrt{\sum_{k=1}^{d} ln^2 \lambda_k(C_i, C_j)} \quad (4)$$

where $\lambda_k(C_i, C_j)$ are the generalized eigenvalues of two covariance matrices C_i and C_j that are defined by the following formula:

$$\lambda_k C_i x_k = C_j x_k \quad (5)$$

where x_k represents the generalized eigenvectors. Note that the generalized eigenvalues λ_k are the solutions of the following equation: $det(\lambda C_i - C_j) = 0$.

It is easy to find the best match if we choose the template with the smallest distance; however, two or more templates can have a similar distance due to noise, while every type of musical symbols have variations of style and shape. It is therefore important to recognize that a large distance does not indicate that two symbols are necessarily different. Distance alone is insufficient to determine whether a symbol s can be classified as a member of the class S_k or not; therefore, we must use a model to transform the similarity to a possibility degree $p_k(s)$ that indicates how likely a symbol s belongs to class S_k. A higher $p_k(s)$ indicates a higher possibility that the symbol s belongs to class S_k. The possibility $p_k(s)$ can be computed from the possibility distribution pd_k that is obtained from the training dataset of each class S_k. The training data contains the images of musical symbols with various kinds of shapes in each class. It contains more than 700 images of musical symbols which are selected from the data of Gamera Project [25].

In our system, the possibility distribution pd_k is described by the following exponential distribution:

$$pd_k(x, \rho) = \rho e^{-\rho x} \quad (6)$$

where x is the distance. The parameter ρ indicates the size of the variant of each class S_k, and it is computed as follows:

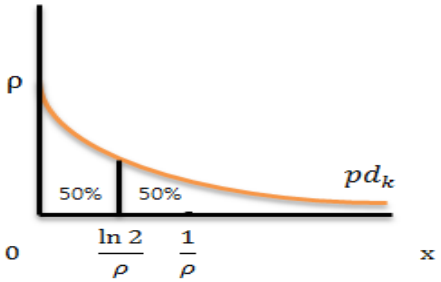

Fig.5. Exponential Distribution

$$\rho = \frac{ln2}{\max(d_k)} \quad (7)$$

where $\max(d_k)$ is the maximum distance of a symbol of the validation dataset that belongs to class S_k.

As described in Figure 5, the region where the distance is smaller than $ln2/\rho$ will cover 50 % of the distribution. We assign this region to correspond with all of the symbols of class S_k in the validation data by using the following to obtain parameter ρ: $\max(d_k) = \frac{\ln 2}{\rho}$.

The value of ρ of every class S_k is different due to the different corresponding variants for the musical style and shape; for example, ρ will be large for the black-note-class symbols since these symbols do not vary much, but it will be smaller in the case of the quarter-rest symbols. It will therefore be "easier" to assign the quarter-rest class than the black-note class.

For the music scores, global information such as the time measures is very significant for the overall musical structure. The time measure is the component that indicates the number of beats that are allowed per bar. Different symbols with the same pitch can have different beats, producing different melodies; for example, a single black note has only 1 beat, a single white note has 2 beats, and a single flag note has 0.5 beats. It is therefore critical to find the time measures so that a decision can be made in the final step. In Figure 6, it is very easy to confuse the white note with the black note, but the time measure indicates that 4 beats must occur in this bar, so the black-note hypothesis will be refused. We generate three hypotheses for the time measures with the covariance-matching process. For each

hypothesis, we compute the prior possibility by

Figure 6. Effect of time-measure information

counting the number of bars in the music score with the same number of beats. The hypothesis that has a maximum value regarding its summation of the possibility degree and the prior possibility will be decided as the final class in terms of the time measures.

The musical syntactic rule is used for the verification of the hypothesis that corresponds to the appropriate class for a symbol. The syntactic rule is as follows: the number of beats per bar must match the time measures. We check this rule for every bar with every hypothesis combination for all of the symbols inside that bar. The objective function is as follows:

$$C_s^k = \begin{cases} \sum_{s=1}^{N} p_k(s) + 1 \; if \; syntactic \; rule \; is \; satisfied \\ \sum_{s=1}^{N} p_k(s) \; if \; syntactic \; rule \; is \; not \; satisfied \end{cases}$$

with Nas the number of symbols inside the bar. The combination that generates the maximum value of C_s^k will be selected as the final class of the symbols inside the bar.

IV. Experimental Results

To evaluate the proposed system, music score images are collected from Korean middle school music text books. The data set includes 50 printed-scanned images and 30 images captured by the camera of a Samsung Galaxy Note 3 as shown in Figure 9. These images are monophonic music scores which do not include multi-notes, multi-beams, multi-flags. More than 4000 symbols belonging to 21 classes are contained in these images. The typical resolution of printed scanned images is 1328 x 1898 while the resolution of the captured images is 2448 x 3264. The template images for musical symbols were collected from Gamera Project [25] as shown as symbols in Table 2(a) and Table 2(b).

We demonstrate the performance of our system by comparing the results to the commercial program Smart Score X2 (Piano Edition) [20], the Deformable LAG proposed by Bui [9], and the SVM classifier of Rebelo [1]. The SVM classifier used the same training data as our proposed model. The SVM toolbox of MATLAB was used for the experiment. Table 1 describes the average accuracies with the printed-scanned images and the captured-scanned images for the proposed method.

Table 1. Average Accuracy

	Printed Images	Captured Images
Smart Score[20]	*96 %*	*65.84 %*
DLAG[9]	*94.3 %*	*82.61 %*
SVM[1]	*96.48 %*	*86.73 %*
Proposed Mtd	*97.44 %*	*96.22 %*

The accuracy of the proposed method is the highest for both datasets, and the difference between the printed-scanned images and the captured-scanned images does not affect the recognition rate of the proposed method as much as it does for Smart Score, SVM, and DLAG. The accuracy of the proposed method with the captured images is decreased by only 1 % from that with the printed images, while the accuracies of the DLAG, SVM, and Smart Score decreased by 12 %, 10 %, and 31 %, respectively. All of the four methods show a similar performance with printed images, but the proposed method is superior in the performance of captured images compared to the other three methods. For the 21 symbol classes, the final recognition error rates of scanned images and captured images by the proposed method are described in Table 2(a) and Table 2(b). The error rates of the proposed system for both scanned images and captured images are less than 5% for most of the important symbols such as single black note, single white note, single flag note, and single beam note; furthermore, the error rates of the other symbols are also smaller than 15%. Generally, the error rates of scanned images and captured images are very similar. The problem, however, is that the time-measurement (no. 22) recognition is not

yet perfect (> 10% of error rate). An incorrect time measurement recognition generates the incorrect number of beats per bar, leading to an inaccurate objective function during the decision-making step.

Unfortunately, a single flag note (no. 5 and 6) of scanned images is the symbol that appears most frequently in these cases of the incorrect time measurement, leading to a higher error rate comparing to the single flag note of captured images.

The decision making step is very useful to boost up the system's performance in the cases of bad binarization or wrong segmentation. Since a white note (no. 3 and 4) can be seen as a black note after binarization, as shown in Figure 7, it is confusing to discriminate a white note from a black note; fortunately, the beat number of these two symbols is different and this can be used to correct the symbols during the decision-making step. The same situation occurs in the cases of the 1/8 rest symbol (no. 17) and the 1/16 rest symbol (no. 18).

A wrong segmentation can also make the symbols become different from their templates as the white note shown in Figure 8. In this case, the white note is merged as one symbol with the touching slur line. The covariance matching process still reserves a white note class among three hypotheses. After that, by counting the number of beat inside this bar, the decision making step picked up the white note hypothesis as the final recognition result.

Table 2 (a). Recognition error rates (1)

No	Symbol	Class Name	Scanned Images	Captured Images
1		*(Single black note)*	*1.14 %*	*1.67 %*
2		*(Single black note)*		
3		*(Single white note)*	*4%*	*4 %*
4		*(Single white note)*		
5		*(Single Flag note)*	*2.1 %*	*1.2 %*
6		*(Single Flag note)*		
7		*(Single Beam note)*		
8		*(Single Beam note)*	*0 %*	*2.1 %*
9		*(Single Beam note)*		
10		*(Single Beam note)*		

Table 2 (b). Recognition error rates (2)

No	Symbol	Class Name	Scanned Images	Captured Images
12		*(Middle Beam note)*	*2 %*	*5 %*
13		*(C clef)*	*0%*	*0 %*
14		*(Flat)*	*4.1 %*	*8.4 %*
15		*(Sharp)*	*8 %*	*9.3 %*
16		*(Natural)*	*9%*	*11.1%*
17		*(1/8 rest)*	*12.1 %*	*11.4 %*
18		*(1/16 rest)*	*9 %*	*0 %*
19		*(Quarter rest)*	*8.3%*	*11 %*
20		*(Whole Note)*	*0 %*	*13.3 %*
21		*(Full & Half rest)*	*0 %*	*0 %*
22		*(Time Measures)*	*12 %*	*10 %*

Table 4 shows the processing time of the proposed model on the Samsung Galaxy Note 3 mobile system. The recognition step that is based on covariance matching only requires 1.5 second for an image of a 1328 x 1898 size. The full system including the pre-processing steps (stave detection, correction, removal, lyric removal, etc.) requires about 3 seconds and this performance is acceptable for a mobile application.

Table 4. Processing time

Image Resolution	Processing time	
	Recognition part	Full system
1328 x 1898	~ 1.5 s	~ 3.2 s

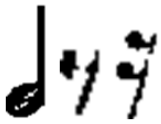

Fig. 7. Binarization result of white note, 1/8 rest and 1/16 rest

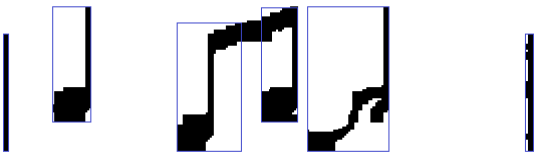

Fig. 8. Segmentation error

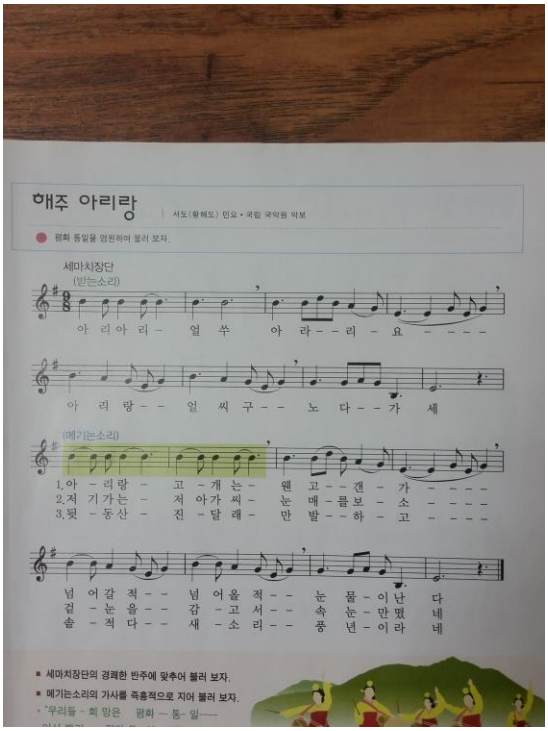

Fig. 9. Music score image

V. Conclusion

In this paper, we propose a novel OMR system based on a LAG and a covariance matching method. The system was designed for recognizing the captured images of monophonic music score. The imprecision and uncertainty from binarization and the segmentation step can be resolved by integrating the structural information with a covariance matching process. The proposed method can correct errors and the final performance is improved. When images are being captured, the covariance-matching algorithm is stable with the variations of a musical symbol's shape and small illumination shifts. The proposed model is also fully automatic and parameter dependencies are limited. The promising results of the current system can be used for a mobile application with an acceptable processing time.

The system is, however, dependent on global information that cannot be perfectly extracted. The system also has a large variant of symbol's shape bias. Symbols with large variant of shape will be easier to assign than symbols with "rigid" shape. In the future, we will investigate a post-processing step so that a case where incorrect structural information is used to make a final decision can be handled. We will also improve the system to deal with the polyphonic music score and the handwritten music score in both scanning and capturing conditions.

REFERENCES

[1] A. Rebelo, G. Capela, and J. S. Cardoso, "Optical recognition of music symbols: A comparative study," *Int. J. Doc. Anal. Recognit.*, vol. 13, no. 1, p. 19-31, Mar. 2010.

[2] D. Pruslin, "Automatic recognition of sheet music," in *Structured Document Image Analysis*, H. Baird, H. Bunke, and K. Yamamoto, Eds. Springer Berlin Heidelberg, 1966.

[3] N. P. Carter and R. A. Bacon, "Automatic recognition of printed music," in *Structured Document Image Analysis*, H. Baird, H. Bunke, and K. Yamamoto, Eds. Springer Berlin Heidelberg, p. 456-465, 1992.

[4] R. Randriamahefa, J.-P. Cocquerez, C. Fluhr, F. Pepin, and S. Philipp, "Printed music recognition," in *Document Analysis and Recognition*, Proceedings of the Second International Conference on, p. 898-901, 1993.

[5] F.Rossant and I.Bloch, "A fuzzy model for optical recognition of musical scores", in *Fuzzy Sets and Systems*, vol. 141, p. 165-201, 2003.

[6] F.Rossant and I.Bloch, “Robust and Adaptive OMR System Including Fuzzy Modeling, Fusion of Musical Rules, and Possible Error Detection”, *EURASIP Journal on Advances in Signal Processing* 2007.

[7] K.T.Reed and J.R.Parker, “Automatic computer recognition of printed music”. *Proc. 13th Int. Conf. Pattern Recognit.*, p. 803-807 (1996).

[8] P. Bellini, B. Ivan, and N. Paolo, "Optical music recognition: Architecture and algorithms," in *Interactive Multimedia Music Technologies*, N. Kia and N. Paolo, Eds. IGI Global, 2008, ch. 5, p. 80-110.

[9] H.N.Bui, “Camera-based Printed Music Score Recognition Using Deformable Line Adjacency Graph”, M.D thesis , Chonnam National University, Gwangju, South Korea.

[10] H.N.Bui, “Camera-based Printed Music Score Recognition Using Deformable Line Adjacency Graph”, M.D thesis , Chonnam National University, Gwangju, South Korea.

[11] O.Tuzel, F.Porikli and P.Meer, “Pedestrian Detection via Classification on Riemannian Manifolds”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.30, No.10, 2008.

[12] H.N Bui, I.S. Na, G.S Lee, H.J. Yang, S.H. Kim, "Boosted Stable Path for Staff-line Detection Using Order Statistic Downscaling and Coarse-to-Fine Technique,"*Proc. 22nd International Conference on Pattern Recognition*, pp. 522-526, Stockholm, Sweden, Aug. 2014.

[13] J.S.Cardoso, A.Capela, A.Rebelo, C.Guedes, “A connected path approach for staff detection on a music score”. In: *Proceedings of the International Conference on Image Processing* (ICIP 2008), p. 1005-1008 (2008).

[14] J.S.Cardoso, A.Capela, A.Rebelo, C.Guedes, J.P. da Costa, “Staff detection with stable paths”. *IEEE Trans. Pattern Anal. Mach. Intell.* 31(6), p. 1134-1139 (2009).

[15] R. Kimmel, N. Kiryati, and A. M. Bruckstein, “Distance maps and weighted distance transform”, Journal of Mathematical Imaging and Vision, Special Issue on Topology and Geometry in Computer Vision, 6:223-233,1996.

[16] R. Andrzej & L. Yongmin, “Learning pairwise image similarities for multi-classification using Kernel Regression Trees.” *Pattern Recognition*, 45(4), 1396-1408, (2012).

[17] L. Chengjun, “Discriminant analysis and similarity measure.” *Pattern Recognition*, 47(1), 359-367, (2014).

[18] K. Takumi, “Kernel-based transition probability toward similarity measure for semi-supervised learning.” *Pattern Recognition,* 47(5), 1994-2010, (2014).

[19] W. Forstner and B. Moonen, “A metric for Covariance Matrices”, Geodesy-The Challenge of the 3rd Millennium

[20] http://www.musitek.com/smartscore-piano.html

[21] B. Gatos, I. Pratikakis, S.J. Perantonis, “Adaptive degraded document image binarization.”, *Pattern Recognition*, 39(3), 317-327, (2006).

[22] H.N Bui, I.S. Na, S.H. Kim, “Staff Line Removal Using Line Adjacency Graph and Staff Line Skeleton for Camera-Based Printed Music Scores.”, *22nd International Conference on Pattern Recognition (ICPR)*, p. 2787-2789, 2014.

[23] L. Pugin, “Optical music recognition of early typographic prints using Hidden Markov models.” In: Proceedings of the International Society for Music, information retrieval, p. 53-56, 2006.

[24] W. Homenda and M. Luckner, “Automatic knowledge acquisition: recognizing music notation with methods of centroids and classifications trees.” In: Proceedings of the international joint conference on neural networks, p. 3382-3388, 2006.

[25] http://gamera.informatik.hsnr.de/addons/musicstaves/index.html.
[26] 도루녹, "Musical Symbols Recognition with Region Proposal Networks and Fast Regional Convolution Neural Network : R-CNNs 과 RPNs 통한 악보인식." 전남대학교: 전자컴퓨터공학과 박사학위 논문 2017. 2

Authors

Luu Ngoc Do

He received the B.S, M.S and Ph.D. degrees from Chonnam National University, South Korea. His main research interests include Data Mining, Pattern Recognition, Machine Learning and Image Processing.

Cong Minh Dinh

He received his B.S. degree in Mathematics & Computer Science from Ho Chi Minh University of Science, Vietnam in 2008, was a software developer in eSilicon Corporation from 2009 until 2012, and received his M. Eng. Degree in Electronics & Computer Engineering at Chonnam National University, Korea in 2015. His research interests are Data Mining and Pattern Recognition.

Hyung-Jeong Yang

She received her B.S., M.S. and Ph.D. degrees from Chonbuk National University, Korea. She was a Post-doc researcher at Carnegie Mellon University, USA. She is currently a professor at Dept. of Electronics and Computer Engineering, Chonnam National University, Gwangju, Korea. Her main research interests include multimedia data mining, pattern recognition, artificial intelligence, e-Learning, and e-Design

Soo-Hyung Kim

He received his B.S. degree in Computer Engineering from Seoul National University in 1986, and his M.S. and Ph.D. degrees in Computer Science from Korea Advanced Institute of Science and Technology in 1988 and 1993, respectively. From 1990 to 1996, he was a senior member of research staff in Multimedia Research Center of Samsung Electronics Co., Korea. Since 1997, he has been a professor in the Department of Computer Science, Chonnam National University, Korea. His research interests are pattern recognition, document image processing, medical image processing, and ubiquitous computing.

Guee-Sang Lee
.
He received his B.S. degree in Electrical Engineering and his M.S. degree in Computer Engineering from Seoul National University, Korea in 1980 and 1982, respectively. He received his Ph.D. degree in Computer Science from Pennsylvania State University in 1991. He is currently a professor of the Department of Electronics and Computer Engineering in Chonnam National University, Korea. His research interests are mainly in the field of image processing, computer vision and video technology.