Multimedia Tools and Applications (2021) 80:11031–11047
https://doi.org/10.1007/s11042-020-10272-2

Binarization of music score with complex background by deep convolutional neural networks

Minh-Trieu Tran[1] · Quang-Nhat Vo[2] · Guee-Sang Lee[1]

Received: 27 May 2020 / Revised: 19 October 2020 / Accepted: 9 December 2020 /
Published online: 3 January 2021

Abstract

Binarization is an important step for most of document analysis systems. Regarding music score images with a complex background, the existence of background clutters with a variety of shapes and colors creates many challenges for the binarization. This paper presents a model for binarization of the complex background music score images by fusion of deep convolutional neural networks. Our model is directly trained from image regions using pixel values as inputs and the binary ground truth as labels. By utilizing the generalization capability of the residual network backbone and useful feature learning ability of dense layer, the proposed network structures can differentiate foreground pixels from background clutters, minimize the possibility of overfitting phenomenon and thus can deal with complex background noises appearing in the music score images. Comparing to traditional algorithms, binary images generated by our method have a cleaner background and better-preserved strokes. The experiments with captured and synthetic music score images show promising results compared to existing methods.

Keywords Binarization · Music score images · Convolutional neural networks · Deep learning · U-net

✉ Guee-Sang Lee
gslee@jnu.ac.kr

Minh-Trieu Tran
tmtvaa@gmail.com

Quang-Nhat Vo
nhat.vo@oulu.fi

[1] Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, South Korea

[2] Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland

1 Introduction

With the development of computer storage technologies and digital cameras, digital archiving and OCR applications are emerging as feasible tools for storing, querying, and processing document information. Among many types of documents, the music score is the one that needs the support of document analysis systems most due to its special musical presentation that creates difficulties for manual digitizing and analyzing. The automatic analysis of music score document images consists of several steps such as binarization, staff line detection and removal, symbol segmentation and recognition. The binarization is the first step and a crucial one in the optical music recognition system. Different from traditional binarization algorithms that are designed for text documents, the binarization of music score needs to consider musical symbols and line strokes. In the case of music scores with a complex background, some foreground-like noises have to be eliminated to create a clean binary map. However, as shown in Fig. 1, the traditional algorithms usually confuse the background noises as foreground symbols due to the use of low-level features of pixels and lack of content knowledge. The image binarization can be considered as a binary classification problem. The presented methods in the literature can be practically divided into two groups: the unsupervised and supervised methods. Most of the available solutions belong to the unsupervised-classification category. The first methods proposed in this group try to assign the image pixels into two

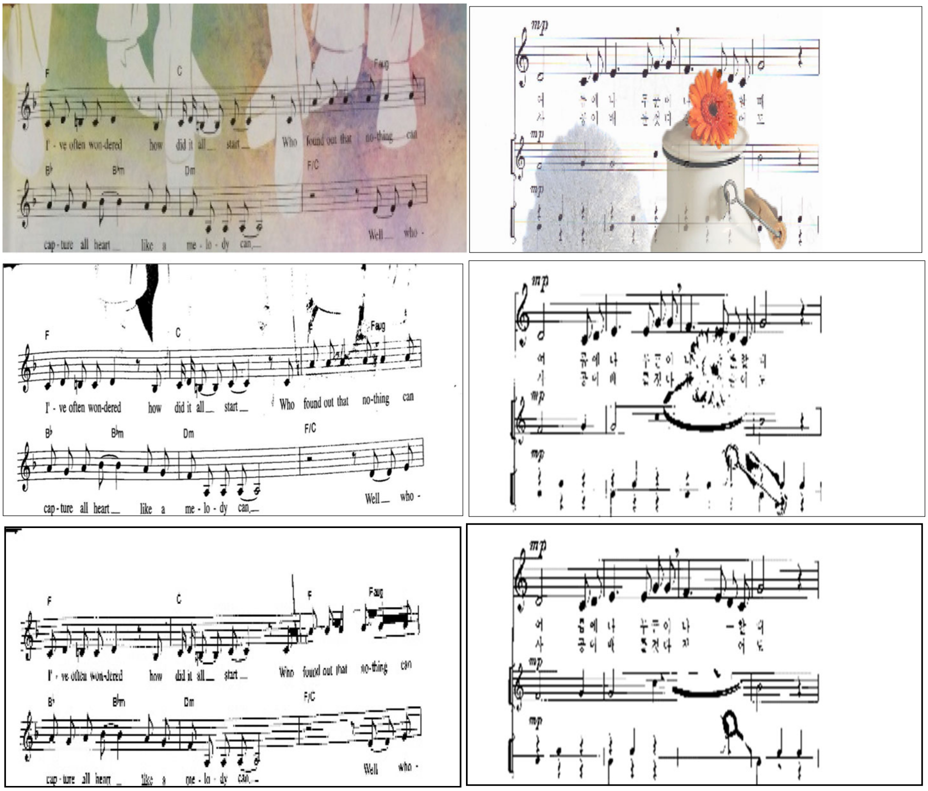
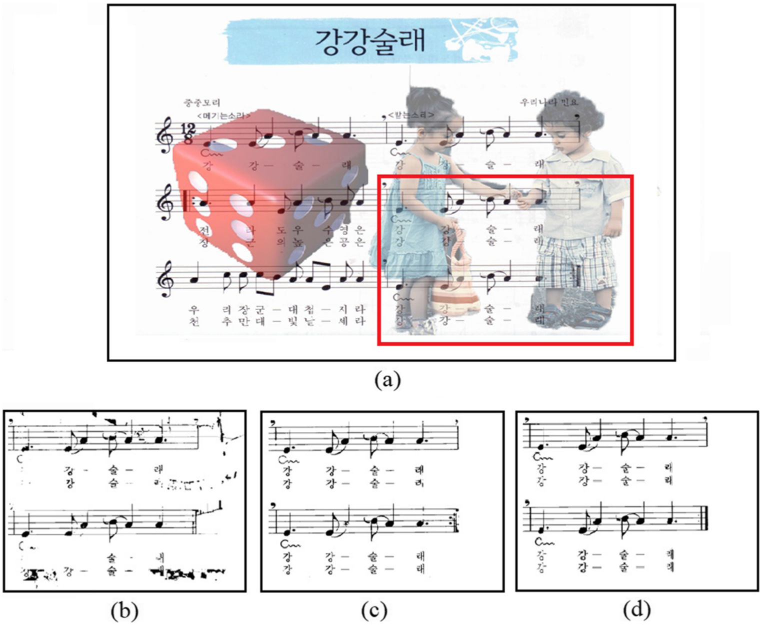

F
C
Faug
I' - ve often won-dered how did it all_ start_ Who found out that no-thing can
B♭
B♭m
Dm
F/C
cap - ture all heart_ like a me - lo - dy can_
Well_ who -
mp

Fig. 1 Demonstration of music score image binarization and challenges. The first row shows the captured and synthetic images. The second and the third rows show the binarization result of Gatos's [4] and Howe's [6] method

classes using some threshold values estimated directly from the image itself. Otsu [11] uses a global threshold value computed form the distribution of all the image pixels. Niblack [10], Sauvola [15] and Gatos [4] compute the local threshold to deal with the non-uniform illumination and color variation. Because these methods only estimate the pixel values, they cannot remove the background pixels that have a similar value to the foreground. Some advance approaches for the classification were proposed. Su et al. [16] apply the Markov-random-field framework for classification of the uncertain pixels into either the background or the foreground based on the already known foreground and background pixels. Howe [6] employs a parameter-turning strategy for the selection of suitable parameter values for each image sample. A Gaussian Mixture Markov Random Field (GMMRF) model [17] was proposed for the binarization of music score images that uses the detected staff lines and background seeds to construct the color distribution of foreground and background. However, this method is difficult in binary images with background similar to the staff line or music symbol. As shown in Fig. 2, the GMMRF model cannot distinguish these patterns in the pant of baby on the right-hand side because it is similar structure to the vertical line of the music score. About supervised learning-based approaches, useful information might be derived from the training-document images. A learning process for the determination of the binarization threshold [1] was introduced. A three-dimensional feature vector is formed by the distribution of the gray values for each image region. The support vector machine (SVM) [3] is trained to classify the feature vectors into four classes corresponding to four decisions of threshold values. Regarding another approach [18], a decision function is learned from different feature types, including the existing and self-developed ones, to directly map an n-dimensions feature vector extracted around a pixel into a binary space. The strong points of the supervised binarization methods are the parameter-free initialization and the absence of the requirement for pre- or post-processing. However, these methods still use low level and hand-draft features



(a)

(b)

(c)

(d)

Fig. 2 The comparison of our binarization result with GMMRF. **a** Input image. **b** GMMRF. **c** Ours. **d** GT

that reduce the robustness of the classification. Our paper proposes a model that combines two deep neural networks. The first network model is deep autoencoder network architecture (DAN) with three skip connections, the second model which we use is the U-net backbone (UB) architecture. The U-net model usually employs in medical imaging segments. However, many challenges appear when we applied U-net to the field of document processing, especially the problem of segmenting images with complex and cluttered backgrounds. The colors of the symbol are similar to the colors of the background. For that reason, we designed a new model based on the U-net architecture with additional modifications to adapt the requirements of this study. On the other hand, a pre-trained residual network was chosen for the backbone structure because it helps model having generalization capability. Different from our previous work [17], we do not need some post and pre-processing steps such as the staff line detection, and our proposed network structure can work on grayscale images while delivers better results. On the other hand, the proposed model can distinguish the foreground and background pixels with similar color values which is the weak point of previous work. The proposed method surpasses state-of-the-art binarization algorithms on captured and synthetic music score images. The contributions of this paper can be summarized as follows. (1) To the best of our knowledge, this is the first time in the literature that a deep supervised network model is used for the binarization of music score images. (2) We propose a sub-model deep autoencoder network architecture (DAN) with three skip connections which have advantages in information flow during the network. (3) We propose a fusion model combine sub-model DAN with the sub-model U-net backbone residual network. The proposed network structures can differentiate foreground pixels from background clutters, and thus can deal with complex background noises appearing in the music score images. (4) The proposed method is evaluated on captured and synthetic images that overcome the other state-of-the-art methods. The rest of this paper is organized as follows: Section 2 describes some related works. Section 3 introduces our proposed approach for the binarization. Section 4 reports the quantitative and qualitative experimental results on benchmarks. Finally, we present a conclusion in Section 5. Our results were quite good, all background details in this area were almost completely removed.

2 Related works

The first binarization methods in the literature focus on finding optimum threshold values for dividing the image pixels into two classes. Otsu [11] proposes the global threshold value that minimizes the variance between the foreground and background distributions. Since the global threshold is sensitive to the local variance, Niblack [10] calculates a pixel-wise threshold in each local region by using the mean and the variance of the gray values in the window. With this approach, the Niblack's method still presents noises if the number of objects in image regions is sparse. To overcome this problem, Sauvola's method [15] defines a new threshold by using speculation on the gray level of foreground and background pixels. However, the binary images of degraded document images still need some enhancements. Gatos's method [4] proposes a combination of existing algorithms. At first, the input image contrast is enhanced by the Weiner filter. Then the Sauvola's method is applied to generate the pre-binary map. Finally, the background noises are suppressed by estimating the neighboring background intensities. About the binarization of music score images, Pinto et al. [13] propose a threshold selection strategy that estimates the staff lines in binary maps at different thresholds. The size of the local window for the local thresholding is also computed from the height of staff lines. The local window size is an important parameter in local binarization methods. Pai et al. [12] propose an adaptive window-size selection method that bases on the image characteristic. On the other hand, Moghaddam and Cheriet [9] introduce a multi-scale binarization framework that is capable of incorporating any local threshold-based binarization method. Another unsupervised binarization approaches are clustering-based and graph-based methods. For example, a clustering-based approach [18] can isolate the foreground from the background by learning an unsupervised model from the pixel features. Su et al. [16] apply the Markov Random Field (MRF) model and a graph-based minimization scheme for classifying uncertain pixels with manually selected background/object pixels. Howe's method [6] introduces an automatic technique for setting parameters of a Markov Random Field model. It depends on a stability heuristic criterion to choose suitable parameter values for individual images. Another binarization method [17] for music score images that combines the Gaussian Mixture Model (GMM) and MRF model is proposed. This method tries to extract the foreground information by modeling the color distribution of detected the staff lines. The label of individual pixels is decided by minimizing an energy function. Different from unsupervised methods, the information that supports the classification of foreground and background in supervised approaches is learned from a set of training-document images. Chou et al. [1] divide an image into sub-areas and construct the decision rules of how to binarize each region. The rules are derived by a training process with the classification of three-dimensional vectors using a support vector machine (SVM). The feature vectors are form from the Otsu threshold, the mean, and standard deviant of each region. The vector space is divided into four classes representing four threshold sections. Wu et al. [18] propose a fully trainable framework for binarization of degraded document images by using the extremely randomized tree. This method introduces two new features, the Logarithm Intensity Percentile (LIP) and the Relative Darkness Index (RDI) which are combined with low-level features such as mean and standard deviation of the entire image intensities, the pixel intensity and its deviation from the Otsu's threshold. The final classification of the n-dimensional feature vectors is then used to predict the class label for all pixels in the document images. Our proposed model is compared with modern networks proposed in recent years such as U-net [14], U-net backbone [19] and Dense U-net [5]. U-net is a very famous model in the field of biomedical image segmentation. This model was developed in 2015, the advantage of this model is the good learning ability featured with the small and medium-size datasets. Dense U-net uses dense layers with the ReLU activation function and skips connection to increase the ability to learn features more effectively for the network. However, the above models exist many problems. When we train the so deep model with small and medium datasets, it is easy to occur overfitting phenomenon. On the other hand, so deep model takes so much time as well as computational hardware requirements, thereby reducing the ability to apply in the real situation. In t he U-net backbone network, the features are studied more fully and completely because of the encoder using residual network [7] architecture with pre-trained weight. Our model is designed that retains the strengths of dense layers, but the activation functions are changed by ELU [2] functions, the network architecture is simplified than dense U-net with only three skip connections between encoder and decoder. Because of the combination with the U-net backbone model, our proposed method could learn features more effectively. The model has the ability in noise reduction and retains more information about notes and staff lines in music score images with complex background.

3 Proposed method

Problem statement In this work, the binary maps of music scores are generated from grayscale images. We focus on the printed music score images that use modern music symbols to indicate the pitches, rhythms, and chords. Every pixel in the original image is assigned to either label 0 (black) or 1 (white), which represents foreground and background, respectively. The foreground of a music score includes pixels of the musical symbols (e.g. Note, Rest, and Chord) and staff lines while the background consists of remaining pixels. The complex background area contains noises such as image patches or non-musical symbols overlapped with the foreground. There are also some foreground-similar features appearing in the background such as edges and high contrast background clutters. The proposed binarization method should be able to separate the true musical information and leave the background noises. In previous works, low-level and handcraft features are still insufficient for distinguishing the foreground from the background; furthermore, most algorithms are designed for some specific datasets with specific characteristics and the adaptation of them to new cases will take time. On the other hand, most of the algorithms are constructed based on simple assumptions of the test images without using the content knowledge. Different from the previous algorithms, the binarization of music score image is addressed in this study through the development of a supervised framework that facilitates the CNN.

Predict model and loss function We formulate the binarization of music score images as a dense prediction problem in which the foreground/background label of a pixel is decided by a likelihood function:

$$P\left(bw_j = y_j | X; \Phi\right); y_j \in \{0, 1\}, \tag{1}$$

where $Y = \{y_j, j = 1,\ldots, |Y|\}$ is the ground truth label of the j^{th} pixel, bw_j is the label of the j^{th} pixel, $X = \{x_j, j = 1,\ldots, |X|\}$ represents the input music score image, and Φ is the prediction model. The function P in Eq. (1) is expected to deliver the highest probability value. Given the image pixels and binary ground truth, our desired prediction model Φ is the one that maximize:

$$\Phi^* = \underset{\Phi}{argmax} \prod_{j=1}^{|X|} P\left(bw_j = y_j | X; \Phi\right), \tag{2}$$

The energy function is computed by a pixel-wise softmax over the final feature map combined with cross entropy loss function. The softmax is defined as:

$$p_k(x) = \frac{\exp(a_k(x))}{\sum_{k'=1}^{K} \exp(a_{k'}(x))}, \tag{3}$$

where $a_k(x)$ denotes the activation function in feature channel k at the pixel position $x \in \Omega$ with $\Omega \subset Z^2$. K is the number of classes, here K is 2 and $p_k(x)$ is approximated maximum function. The cross-entropy then penalizes at each position the deviation of $p_{l(x)}(x)$ from 1 using:

$$E = \sum_{x \in \Omega} w(x) \log\left(p_{l(x)}(x)\right), \tag{4}$$

where $l : \Omega \rightarrow \{1, 2\}$ is the true label of each pixel and $w : \Omega \rightarrow \mathbb{R}$ is a weight map that we introduced to give some pixels more importance in the training. During training process, we used binary cross entropy loss funtion with ADAM optimizer. The binary cross entropy is defined mathematically as:

$$BCE(t,p) = -\Big(t{*}\log(p) + (1{-}t){*}(\log(1{-}p)), \tag{5}$$

where t is a correct target and p is a predicted value. Entropy is measure of uncertainty in a certain distribution and cross entropy s the value representing the uncertainty between the target distribution and the predicted distribution.

Model We propose a binarization system that employs a fusion of deep convolutional neural networks for the extraction of the musical symbols from the music score images. Figure 3 shown our proposed architecture. The modification of dense layers is detailed as in Fig. 4. Each network is learned independently using image patches as input and binary maps as ground truth. The design of our proposed binarization model allows the networks to predict effectively the foreground maps. Our proposed model learned useful features from the images and it has generalization capability. In recent years, the advantage of the backbone model has been proved effective. Because of pre-trained weight which was trained with a large dataset such as Imagenet, the generalization capability of the backbone network is improved significantly. Although it is possible that the distributions in our evaluation data set are not included in the Imagenet dataset, pretrained networks are very useful in getting the first layers of the network. In fact, the first convolution classes are the classes that define the specific shapes of edges, lines, oblique lines, etc. and these are the most difficult classes to focus on during model training. The model can learn useful features more effectively while the spending time for training is reduced. In the [8] study, full pre-activation architecture is chosen for the most optimal because the error rate generated is lowest, minimize the possibility of overfitting phenomenon. We use the variant residual block with pre-activation architecture as the basis for the backbone residual network in our network model. The block architecture is described in detail as Fig. 5. A detail of the decoder block is shown in Fig. 6. The visualization of feature maps in the first few layers are shown in Fig. 7. On another hand, we propose a deep autoencoder network with three skip connections. The encoder of our proposed model used a modification of dense layers to enhance the information flow of the useful feature during the

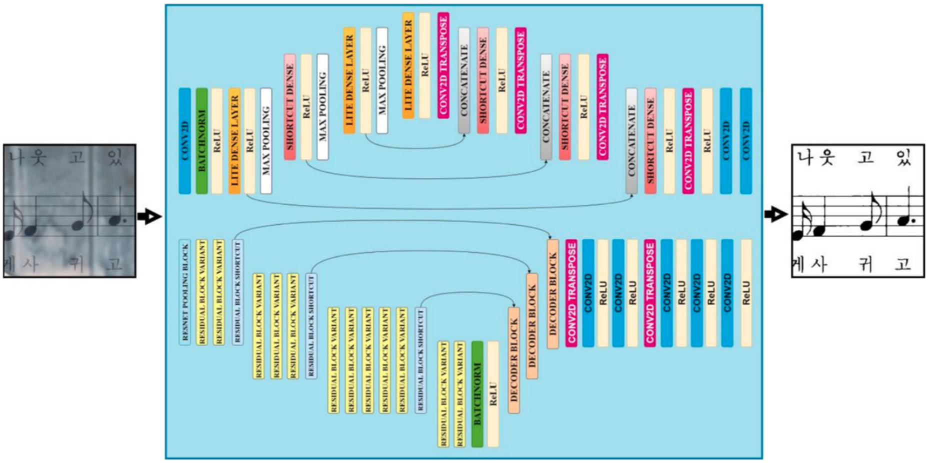


Fig. 3 Our proposed structure

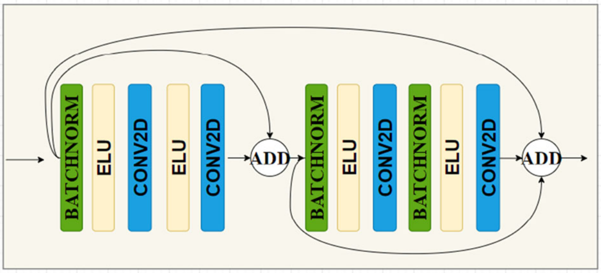


Fig. 4 Modification of dense layers

training process. Therefore, featured information is better preserved. The results of the second model are better in synthetic images. However, for captured images, the separate use of the two models does not produce the same good results as when they are combined. We take advantage of the ELU activation function in this architecture. The ELU function has been shown to make the learning model more efficient, saving model training time more than the ReLU function [2].

Network architecture Each sub-model of our fusion model is composed of several groups of convolutional layers and activation function and skip connection. We also choose to use 2 × 2 pooling windows in all pooling layers. Training data for our model are image patches and corresponding ground truth binary maps sampled from collected music score images. All image patches are converted to grayscale. Two models are trained independently to predict the foreground maps at different feature levels. The cross-entropy loss is computed at side-output layers for back-propagation. About the testing, to be compatible with the training phase, we divided entire images into patches. The patches of size 256 by 256 were created from entire images. This paper proposes a method of combining convolutional neural network models for the task of binarization of complex background music images. In the first model, the deep learning model was designed based on the U-net network, dense layers were integrated into the encoder as well as the decoder of the model. The main purpose of the proposed dense layers is to create the ability to learn the characteristics of the object more deeply, the model can distinguish the pixel value is information or noise need to remove. The structure of dense layers at the encoder block and decoder block is mentioned in Figs. 4 and 5. The next model used to combine is U-net architecture with the backbone. In this work, we use cross-entropy loss with dice loss in the training phase. The advantage of this model is that the weights of the encoder are pre-trained with the large Imagenet data set which helps the model having generalization capability as well as the ability to prevent overfitting phenomenon. The residual network based on variant residual blocks was chosen because of its advantages in learning useful features, as well as the ability to minimize the risk of overfitting [7]. We use full pre-

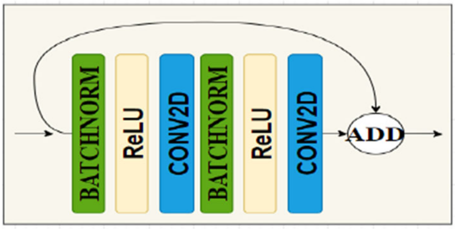
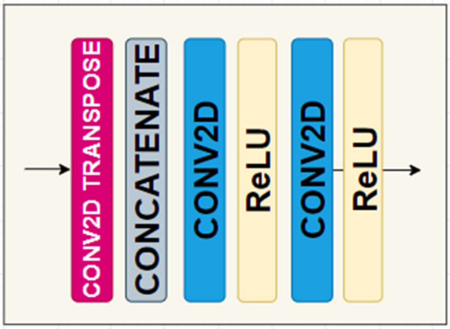


Fig. 5 Variant residual block structure

Fig. 6 Decoder block structure



activation to establish the residual network at the backbone of U-net because the classification error rate of its is the lowest [8] Figs 6 and 7.

Binarization Two foreground maps are predicted for each image patch by our architecture. The next step is to generate two binary maps by two sub-networks *DDLU* and *UB*, for the predicted foreground maps by applying the threshold *T*:

$$bw_j = \begin{cases} 0 & \text{if } f_j \geq T \\ 1 & otherwise \end{cases}, \tag{6}$$

where f_j is the predicted value of j^{th} pixels in the foreground map. To select the optimal threshold value *T*, we analyze *T* on the predicted maps of the randomly selected training images. The *F-measure* of binary images generated at different values of *T* is estimated to select the optimum one. As shown in Fig. 8, by selecting $T = 0.976$, the highest *F-measure* value is given. After obtaining the binary map of each prediction level, the final binary map is composed as shown in Eq. (2). From processed image patches, the complete foreground map of the tested image is recovered.

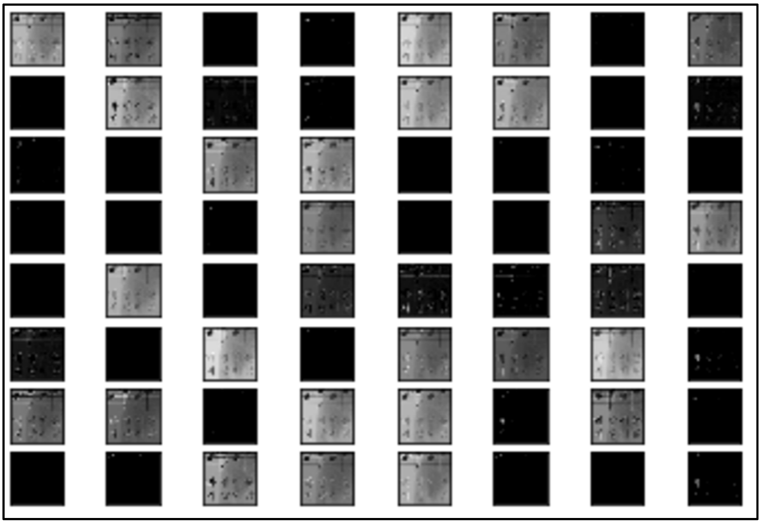


Fig. 7 Feature maps visuallization in first few layers

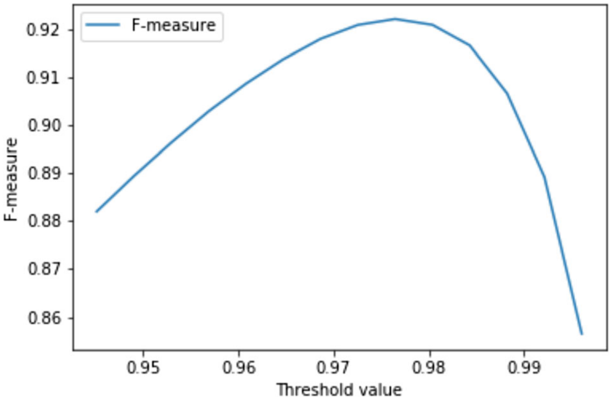


Fig. 8 Distribution of F-measure estimated on binarization of predicted maps of random selected training images with different T values

Calculate the combined weight We propose linear equations containing the number of combinations between two predicted maps of the two networks. Alpha is the factor for the DAN network and beta is the factor for the U-net backbone network. The sum of the two weights is equal to 1. Table 1 presents the empirical values when choosing the values of combined weight. As observed, with an alpha value of 0.8 and a beta of 0.2, the *F-measure* value has the best result. This *F-measure* value is calculated on a dataset including both captured music score images and synthetic music score images.

$$O^p = \alpha O^{DAN} + \beta O^{UB}, \quad \alpha + \beta = 1 \tag{7}$$

where O^p is a prediction of output, O^{DAN} is a prediction of DAN model output, O^{UB} is a prediction of UB model output, α and β are weights for combination.

4 Experimental results

4.1 Datasets and evaluation metrics

There are two sets of music score images that are used in this work for making the training data. Test images and training data are gathered from captured and synthetic music score

Table 1 Empirical values when choosing the values of combined weight

α	β	*precision*	*recall*	*specificity*	*F-measure*
0.1	0.9	0.8074	0.9720	0.9844	0.8788
0.2	0.8	0.8211	0.9709	0.9860	0.8866
0.3	0.7	0.8358	0.9694	0.9876	0.8946
0.4	0.6	0.8510	0.9672	0.9891	0.9024
0.5	0.5	0.8687	0.9636	0.9908	0.9108
0.6	0.4	0.8866	0.9582	0.9924	0.9181
0.7	0.3	0.9067	0.9477	0.9942	0.9238
0.8	0.2	0.9303	0.9274	0.9961	0.9258
0.9	0.1	0.9531	0.8552	0.9978	0.8985

images. The binary map ground truth of each music score image is created manually by ourselves. Figure 9 presents some samples of our dataset. Captured images: This set contains captured music score images with free-form color background and illumination. The average resolution is about 2448 by 3264. In total, we have collected 80 images for creating the test data. Synthetic images: Images in this set are printed music score images with background blended with some graphical figures. In total, we have collected 20 images for creating the test data. The average resolution is about 2840 by 3570.

For the evaluation, we calculate the well-known measures such as *precision, recall, F-measure, and specificity.*

$$F\text{−}measure = 2\frac{recall \times precision}{recall + precision}, \tag{8}$$

$$recall = \frac{TP}{TP + FN}, \tag{9}$$

$$precision = \frac{TP}{TP + FP} \quad , \tag{10}$$

$$specificity = \frac{TN}{FP + TN}, \tag{11}$$

where *TP*, *FP*, *TN*, and *FN* are the true-positive value, the false-positive value, the true-negative value, and the false-negative value, respectively.

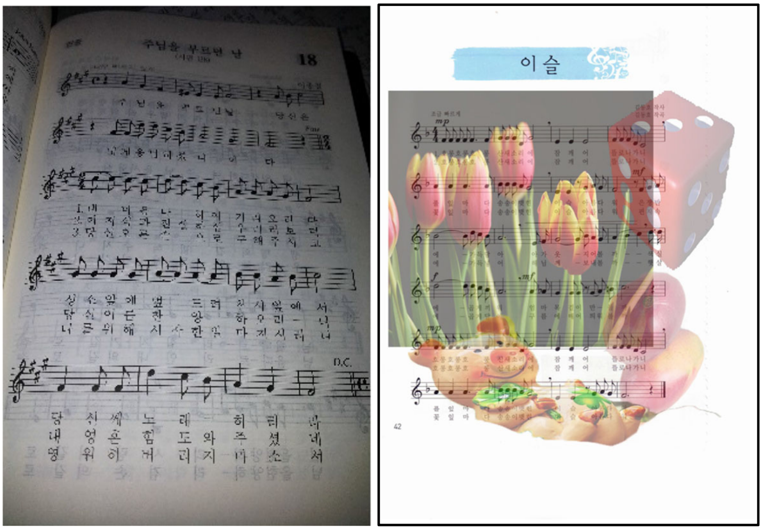


Fig. 9 Samples of our collected music score images. The datasets include captured images and synthetic images

4.2 Implementation

Data preparation Collected music score images in both datasets are used for the training. All of the captured images are employed for creating the training image patches. We employed 100 images include 80 captured images and 20 synthetic images for the testing. We also do the augmentation by rotation patches and binary maps with the scale factors 0.2, width shifting, height shifting and resizing patches and binary maps with the scale factors 0.05. Overall, approximately 100,000 training image patches were created. Figure 10 displays some image-patch samples from the training data.

Parameter and setting We train our DSNs over created image patches. Two sub-models are learned independently. During the process of learning our model, we set the training epoch to 30. We set base learning rates of 10–7. Our network is trained and tested with Keras framework. The proposed model runs on a PC platform with a 3.6GHz core i7 7700 CPU, 24 GB memory, and a single NVIDIA GTX 1080.

4.3 Results

The performance of the proposed method is demonstrated on both captured images and synthetic images. The testing is performed on collected images of our previous work [17] and 20 created synthetic images. We compare our method with the state-of-the-art algorithms including Gatos et al.'s method [4], Howe's method [6], U-net [14], Dense U-net [5], U-net backbone [19], and GMMRF method [17]. The threshold T = 0.97 is applied to the predicted maps of DAN and UB for creating the binary maps. Tables 2 and 3 describes the quantitative evaluation in terms of *precision*, *recall*, *specification* and *F-measure*. Besides the final results

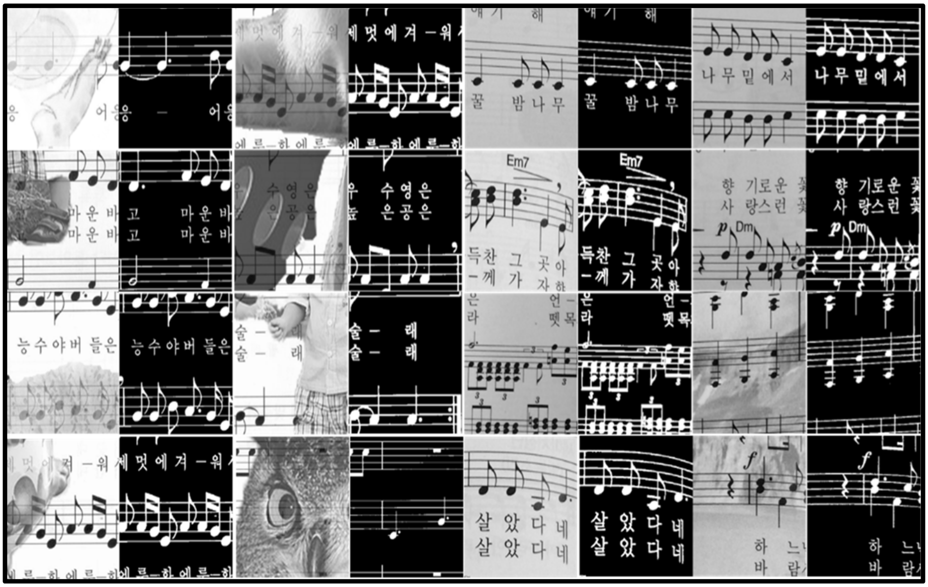


Fig. 10 Samples of our generated training image patches and ground truth binary maps

Table 2 Quantitative evaluation for the binarization of captured music score images

Methods	*precision*	*recall*	*specificity*	*F-measure*
Gatos	0.8920	0.9149	0.9917	0.9032
Howe	0.8272	0.9541	0.9810	0.8858
GMMRF	0.9763	0.8833	0.9983	0.9227
U-net	0.7253	0.9733	0.9720	0.8244
Dense U-net	0.8534	0.9205	0.9908	0.8806
UB	0.6227	0.9910	0.9539	0.7596
DAN	0.8767	0.9056	0.9914	0.8856
Proposed	0.9214	0.9321	0.9954	0.9235

of our approach, the intermediate results of the individual sub-model (DAN and UB) are also presented. It is clearly shown that our binarization model can extract the foreground pixels and remove the background noises better than other methods. By using the GPU, the average processing time of our system is around 7 s for one image of size 2840 by 3570. About the training time, it takes around 50 h for a training DAN network. The training time can be reduced with a stronger system and two sub-models can be trained independently.

Captured music score images Test images in this set contain free-form background shapes that were captured by the smartphone camera. The quantitative evaluation is presented in Table 2. Our method performs the best in terms of *F-measure* which denotes the balance in the preserve of the foreground and the removal of background. Our structures allow us to eliminate the background noises and refine the foreground pixels, and hence, improve the *precision* score and *F-measure* of the final results. Although the GMMRF has the highest *precision* value, its *recall* value is lower than our method. As displayed in Fig. 12, our method and the GMMRF method generate the best visual quality in binary images. The evaluation demonstrates the robustness of the proposed model in the binarization of music score images.

Synthetic music score images To demonstrate the power of our binarization model in case of an extremely complex background, the proposed method is tested with synthetic images. These images are more challenging than the captured ones and include the weak foreground information and dominant background clutters. The binary images of evaluated methods are presented in Figs. 11 and 12. As we can see, the other methods either fail to remove the background or corrupt the staff lines and vertical bars. The GMMRF method also leaves some noises in the background due to the similarity of the foreground and background color. By using high-level features, our approach can predict the correct foreground pixels and delivers

Table 3 Quantitative evaluation for the binarization of synthetic music score images

Methods	*precision*	*recall*	*specificity*	*F-measure*
Gatos	0.6698	0.9008	0.9785	0.7557
Howe	0.6444	0.9183	0.9786	0.7525
GMMRF	0.8622	0.8494	0.9942	0.8520
U-net	0.6867	0.9639	0.9821	0.7996
Dense U-net	0.8819	0.9047	0.9952	0.8914
UB	0.6588	0.9832	0.9800	0.7869
DAN	0.9225	0.9284	0.9969	0.9245
Proposed	0.9658	0.9084	0.9988	0.9352

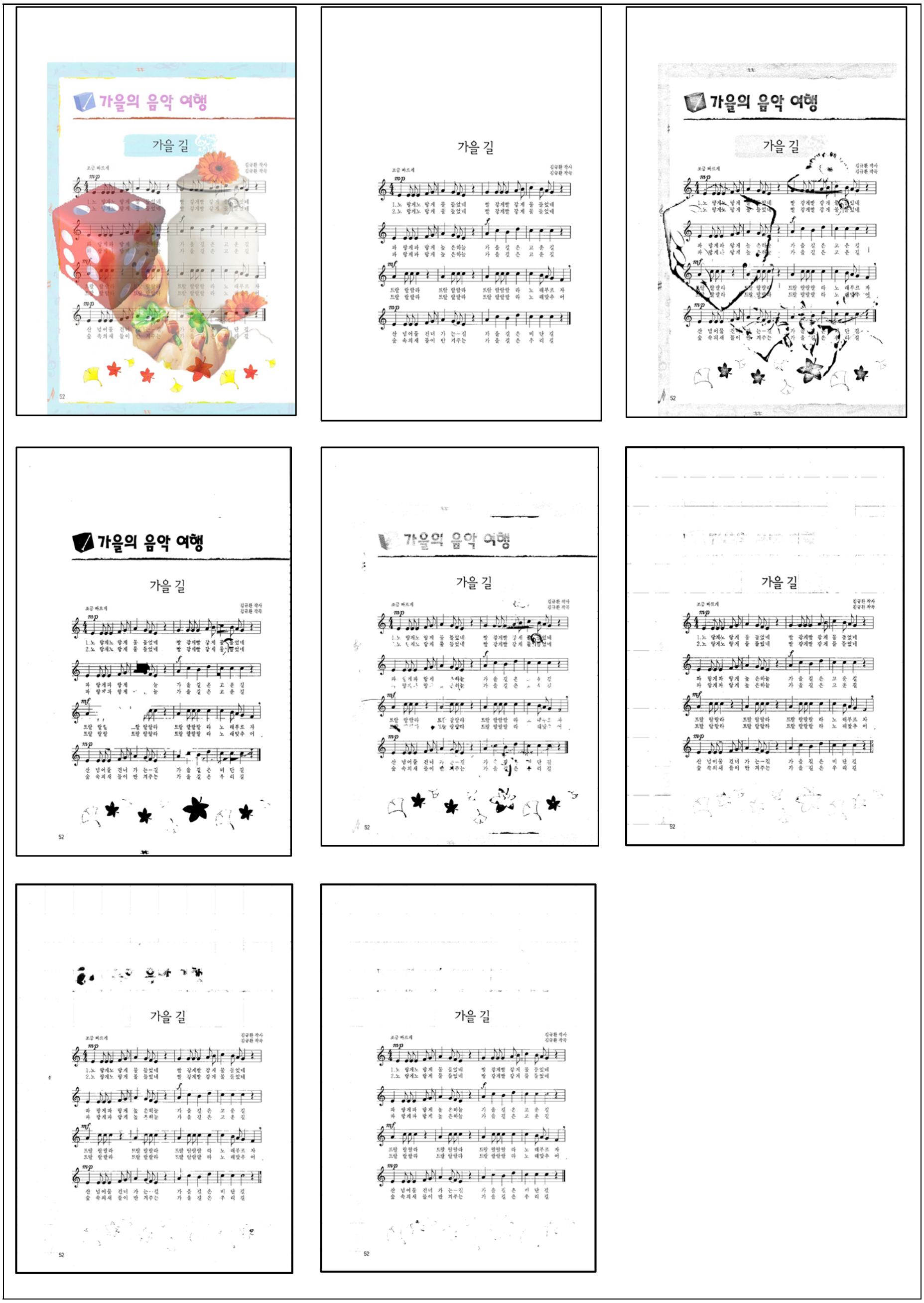


Fig. 11 The binarization results of our method for a synthetic music score image compared with the others. From the left to the right, top to bottom: the original image, the ground truth, Gatos's method, Howe's method, GMMRF method, [5, 14] and proposed fusion model

the best visual quality on sample images. Table 3 shows the quantitative evaluation that demonstrates the advantages of our methods in terms of *precision* and *F-measure*. The results of each sub-model, DAN and UB are also comparable to the results of state-of-the-art

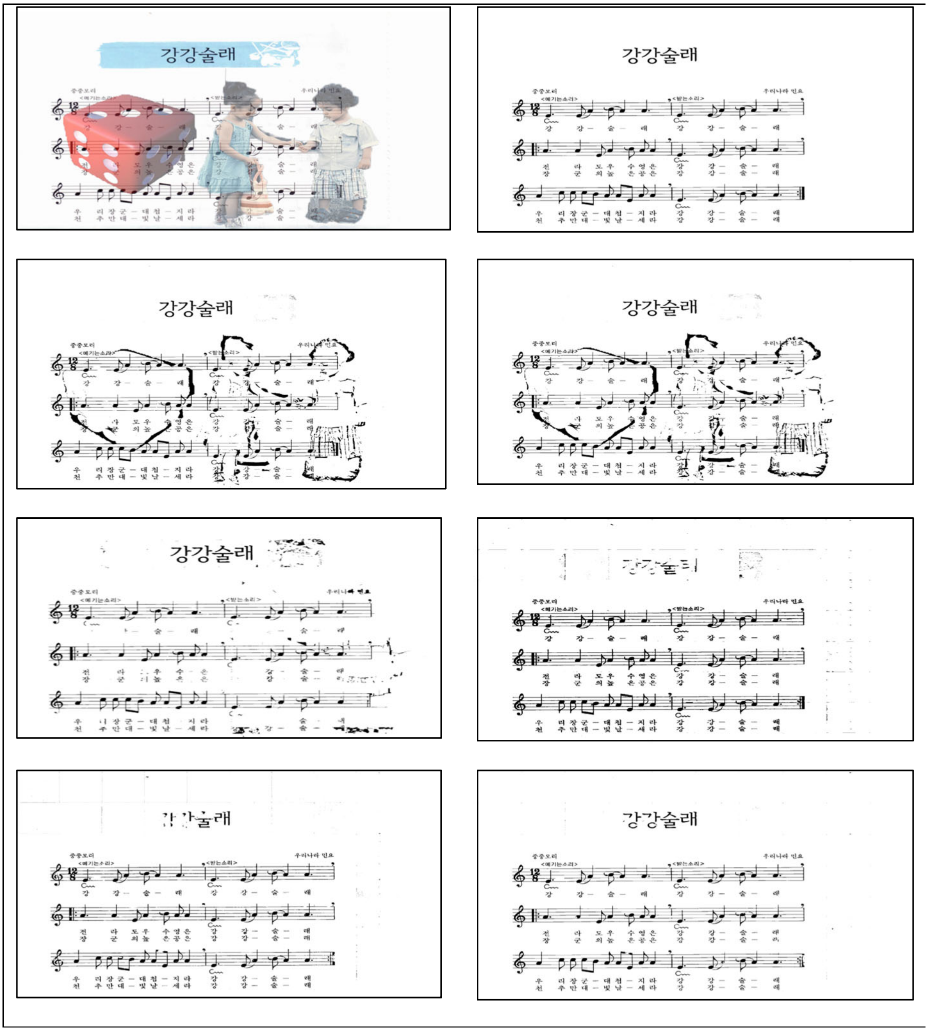


Fig. 12 The binarization results of our method for a synthetic music score image compared with the others. From the left to the right, top to bottom: the original image, the ground truth, Gatos's method, Howe's method, GMMRF method, [5, 14] and proposed fusion model

algorithms. Similar to the case of captured images, the integration of predictions of two network structures can increase the *precision* score with a little reduction of *recall* values and improve the *F-measure*.

5 Conclusions

In this paper, a novel supervised-binarization framework for the music score images with a complex background is proposed. The binarization model is based on a fusion of deep convolutional neural networks. We have shown that our model learned with created training samples can differentiate musical notations from background noises efficiently. The evaluation of two collected datasets demonstrates that the proposed method outperforms the state-of-the-art binarization algorithms. Our method can preserve the foreground information better and can provide excellent visual quality. Not only focusing on the music score images, but we also consider the application of our approach for binarization of other types of documents, such as historical documents and paychecks.

Acknowledgements This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3B05049058 & NRF-2020R1A4A1019191).

References

1. Chou CH, Lin WH, Chang F (2010) A binarization method with learning-built rules for document images produced by cameras. Pattern Recog 43:1518–1530
2. Clevert D, Thomas U, Hochreiter S (2016) Fast and accurate deep network learning by exponential linear units (ELUs). In: Proceedings of the International Conference on Learning Representations. arXiv preprint arXiv:1511.07289
3. Cortes C, Vapnik V (1995) Support-vector network. Mach Learn 20:273–297
4. Gatos B, Pratikakis I, Perantonis SJ (2004) An adaptive binarization technique for low quality historical documents. Lecture Notes Comput Sci: Doc Anal Sys VI:102–113
5. Guan S, Khan A, Sikdar S, Chitnis P (2020) Fully dense UNet for 2D sparse Photoacoustic tomography artifact removal. IEEE J Biomed Health Informatics 24:568–576
6. Howe NR (2012) Document binarization with automatic parameter tuning. Int J Doc Anal Recog 16:247–258
7. Kaiming H, Xiangyu Z, Shaoqing R, Jian S (2016) Deep residual learning for image recognition. In: Proceedings of the International Computer Vision and Pattern Recognition. IEEE, Las Vegas, pp 770–778
8. Kaiming H, Xiangyu Z, Shaoqing R, Jian S (2016) Identity mappings in deep residual networks. In: Proceedings of the European Conference on Computer Vision. Springer, Cham, pp 630–645
9. Moghaddam RF, Cheriet M (2010) A multi-scale framework for adaptive binarization of degraded document images. Pattern Recog 43:2186–2198
10. Niblack W (1986) An introduction to digital image processing. Strandberg Publishing Company:115–116
11. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern Syst 9:62–66
12. Pai YT, Chang YF, Ruan SJ (2010) Adaptive thresholding algorithm: efficient computation technique based on intelligent block detection for degraded document images. Pattern Recog 43:3177–3187
13. Pinto T, Rebelo A, Giraldi G, Cardoso JS (2011) Music score binarization based on domain knowledge. Pattern Recog Image Anal, Lect Notes Comput Sci 6669:700–708
14. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. Med Image Comput Comput Assist Interv, Lect Notes Comput Sci 9351:234–241
15. Sauvola J, Pietikinen M (2000) Adaptive document image binarization. Pattern Recogn 33:225–236

16. Su B, Lu S, Tan CL (2012) A learning framework for degraded document image binarization using Markov random field. In: Proceedings of the international conference on pattern recognition. IEEE, Tsukuba, pp 3200–3203
17. Vo QN, Kim SH, Yang HJ, Lee GS (2016) An MRF model for binarization of music scores with complex background. Pattern Recog lett 69:88–95
18. Wu Y, Natarajan P, Rawls S, AbdAlmageed W (2016) Learning document image binarization from data. In: Proceedings of the international conference on image processing. IEEE, Phoenix, pp 3763–3767
19. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J (2019) UNet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans Med Imaging 39:1856–1867