Patch-Based Convolutional Neural Network for the Writer Classification Problem in Music Score Images

Leandro Takeshi Hattori[1], Matheus Gutoski[2]
Nelson Marcelo Romero Aquino[3], and Heitor Silvério Lopes[4]

Federal University of Technology - Paraná
Av. Sete de Setembro, 3165 - Rebouças CEP 80230-901
[1] lthattori@gmail.com
[2] matheusgutoski@gmail.com
[3] nmarceloromero@gmail.com
[4] hslopes@utfpr.edu.br

Abstract The Writer Identification Problem has been largely studied in the field of image processing. Music score writer identification is a particular type of the problem that requires identifying the writer of a music score, which is a complex task even for musicologists. Addressing this issue, this paper presents a novel Deep Learning approach based on a Convolutional Neural Network (CNN) for classifying music score images according to their writer. The classification is accomplished by dividing a music score image into patches that are fed to the CNN, which provides classification results for each patch. A voting system is then applied to obtain the final prediction of the model. This approach allows to learn local features of each music score in order to improve the final classification result. Results show that the proposed approach allows to obtain satisfactory results for the dataset used in this work, reaching 84%, 94% and 98% for the top-1, top-3 and top-5 accuracies, respectively.

Keywords: Music Scores Classification, Handwritten Classification, Deep Learning, Convolutional Neural Networks

1 Introduction

The understanding of music scores is an area of growing interest of Document Image Analysis and Recognition (DIAR) [1]. Among the many problems of DIAR, the Writer Classification Problem (WCP) has been an area of growing interest in the Computer Science community [2, 3, 4]. Identifying the writer of a music score is a complex and time consuming task even for musicologists. One of the main reasons for the complexity of the task is the large amount of music writers. Hence, Computer Science plays an important role in developing approaches to automatically perform tasks such as the WCP.

An alternative way to work with the inherent complexity of writer identification is to use traditional Machine Learning (ML) techniques, which have shown great effectiveness at finding patterns in data [5] [6]. More specifically, Deep Learning (DL) methods such as Convolutional Neural Networks (CNNs) have recently achieved state-of-the-art performance in many fields [7], including handwritten document image classification tasks [2] [4].

The WCP can be approached as a multi-class classification problem. In WCP, the goal is to identify the writer of a handwritten document. There are two different approaches to this problem: on-line and off-line classification. In on-line classification, the goal is to predict the writer in real time by using temporal and spatial information. In off-line classification, the only information available is a static image of the written document.

A common step prior to the music score WCP is the removal of staff lines. Successfully segmenting the symbols in a music score can lead to a better classification performance [8]. Due to its importance, many algorithms have been proposed for the staff line removal problem [9].

This work presents an approach to predict the writer of staff-less music score images from the CVC-MUSCIMA dataset in an off-line manner. This is accomplished by using a CNN. We also propose a patch extraction strategy that uses a sliding window to better capture local information of the music scores. At last, a voting system is used for classification, where each patch of the original image is classified separately, and the class with most votes wins.

This paper is organized as follows: Section 2 introduces some important literature methods for the WCP. In Section 3, we present a review of Convolutional Neural Networks. Next, in Section 4, we present the methodology for patch extraction and music score classification. Next, in Sections 5 and 6, the computational experiments and results are detailed. Finally, in the last Section 7, discussion about results, conclusions and future directions are pointed out.

2 Related Work

Many methods have been proposed to tackle the writer identification problem in document images. In [10], the Histogram of Oriented Gradients (HOG) was applied to Arabic written document images. In [11], a bagged discrete cosine transform (BDCT) descriptor was used to identify the authorship of English written document images.

In the writer identification problem applied to music score images, two recent works are worth citing. In [12], a Hidden Markov Model (HMM) was used with a Blurred Shape Model (BSM) descriptor, applied to the CVC-MUSCIMA dataset without staff removal. In [13], Hinge features were used in a small dataset with 88 music score samples. The work presents results with the global analysis of the image. However, the author suggests an improvement applying the Hinge feature in small fragments of the music score images.

Deep Learning (DL) methods have been highlighted in recent years, with applications in WCP [7]. In the work presented in [2], a Convolutional Neural

Network (CNN) was used to perform the classification of images containing writings from different nationalities. Another work, presented by [4], used a CNN to recognize the authorship of signatures in images.

In this context, several works regarding WPC [2, 3, 4] presented approaches based on the use of hand-designed feature extractors. Some of them were applied to solve problems related to music score documents [12, 13]. In contrast, this work proposes the use of a CNN to learn the feature extractor and the classifier jointly, which is an approach that has already given satisfactory results in problems belonging to a wide variety of areas [7] [2] [4]. Another approach that has been explored in recent works, similar to those presented by [10] and [13], is based on the analysis of local characteristics of images by extracting patches. The patch extraction strategy also presented good results on medical image analysis [14]. Since this method has also been proved to provide satisfactory results, we propose the classification of individual patches of a music score and perform a voting system to determine the author of the piece.

Similarly to [12], this work uses the CVC-MUSCIMA dataset. However, we propose a novel classification strategy for the dataset and the use of the hold-out validation method, which consists in dividing the dataset into train and test sets. Another differentiating factor is that we use the music scores without staff lines, since staff-less images provide better classification performance [9]. Therefore, a direct performance comparison between our method and the one presented in [12] is inappropriate.

3 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a special type of feed-forward neural network that can automatically learn high-level representations from raw images. This process is accomplished by means of a multistage trainable network, in which each stage usually contains a convolutional layer followed by a non-linearity layer and a pooling layer [15]. Just after the layers responsible for extracting features, Fully Connected (FC) layers are added to perform the classification task itself. This architecture allows the network to learn high-level features and the classifier for a specific problem.

Along with the growth in popularity, CNNs have also grown in size and complexity [16]. Deeper architectures can learn more complex representations, which may lead to better classification performance. However, CNN are also susceptible to well known problems, which can be minimized by using certain techniques.

Proposed by [17], nowadays the Rectified Linear Units (ReLU) is considered to be the most popular non-linear activation function. It is a half-wave rectifier $f(x) = max(x, 0)$ (where x is the input to a neuron). The advantage of the ReLU, compared with other logistic functions, such as sigmoid and hyperbolic tangent functions, is that it learns much faster in multi-layer networks, allowing the training of a deep supervised network [7].

The Local Response Normalization (LRN) [18] is a method that amplifies the excited neuron while fading the surrounding neurons. This method performs well combined with ReLu activation, given that ReLu has unbounded activations and LRN normalizes them.

Pooling layers are a standard layer in CNN models. This layer down-samples the data representation in order to reduce the amount of parameters of the model and reduce the computational time required to train the network.

A common problem in deep architectures is overfitting. The term is used to describe the lack of generalization capacity of the network. The problem happens when the network gets overly adapted to the training set. An overfit CNN shows good performance on the training phase. However, images not contained in this set are likely to be misclassified. There are several techniques to minimize the effects of overfitting in deep neural networks, such as dropout [19] and regularization [20].

Dropout is a technique that randomly eliminates a certain number of neurons in a specific layer with a priorly defined probability, creating a noise effect. This process prevents the co-adaptation of neurons, meaning that one unit is not dependent on the presence of another unit to make a prediction [21]. In general, studies report that this technique improves the CNN performance and reduces overfitting [22].

L1 and L2 regularization are techniques that try to force the network weight values to be as small as possible by adding a term to the error function. L1 regularization attempts to minimize the sum of the absolute weight values, whilst L2 regularization tries to minimize the squared sum of the weight values [20]. Regularization reduces overfitting by forcing the network to use every neuron instead of a small group of neurons, thus increasing the generalization capacity of the network.

4 Methodology

4.1 Data preprocessing

In the image WCP, local features can wield useful information for classifiers. In order to capture this local information, we propose a patch extraction process, which breaks the original image into smaller parts. Moreover, this process also provides a way to augment the data, which benefits DL methods.

For the patch extraction process, we employ a sliding window of 200×200 pixels with a stride of 100 pixels. We select patches with more than 5% of relevant information and discard the rest (in our case, patches that contain mostly black pixels). Afterwards, images are reshaped to 100×100 pixels. The process is presented in the Figure 1.

4.2 Patch-Based Convolutional Neural Network

In this section we propose the method for music score writer classification using a Patch-Based CNN. The CNN architecture used in this work is based on Levis

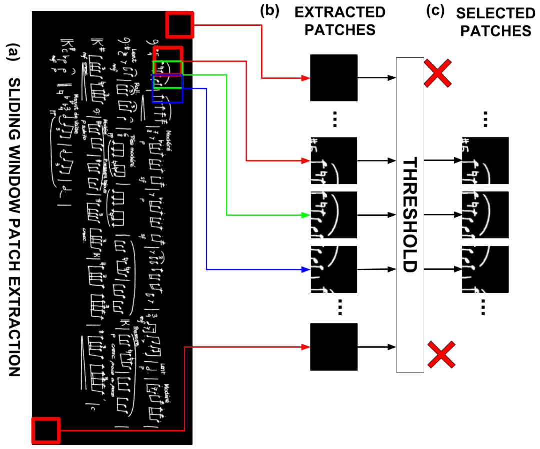


Figure 1. Overview of the extraction and selection of patches in the preprocessing step.

[23] work, as shown in Figure 2. It contains three convolutional layers (conv1, conv2 and conv3) after the input layer, each followed by a max pooling and LRN layers (maxpool1, maxpool2 and maxpool3). Three FC layers: two layers with 512 neurons each (dense1 and dense2), and an output layer with the number of neurons required by the problem. There are two dropout layers: the first between the dense1 and dense2 layers, and the second between the dense2 and out layer, both with 50% drop probability. Then, a *softmax* activation function determines the predicted class. In this work, we use the Glorot initialization method for initializing the weights of the network and the ReLU activation for all convolutional and FC layers (see Section 3). We also adopt the L2 regularization (see Section 3).

The cost function used to train the Patch-Based CNN is the cross-entropy between the predicted ($p_{i,j}$) and target ($t_{i,j}$) writer classes, as presented in Equation 1, which is the loss function (L_i) for multi-class problems with softmax output.

$$L_i = -\sum_j t_{i,j} \log(p_{i,j}) \tag{1}$$

Gradient optimization approaches are used in order to minimize the gradient of the network and to control the value of the learning rate (η) which, in turn, determines the size of the steps that are taken towards the minimum error based on the cost function. Therefore, η defines the velocity at which the network reaches the minimum during the training process. The Adam optimizer approach was used, which is powerful and well-suited to a large amount of optimization problems in DL applications [24].

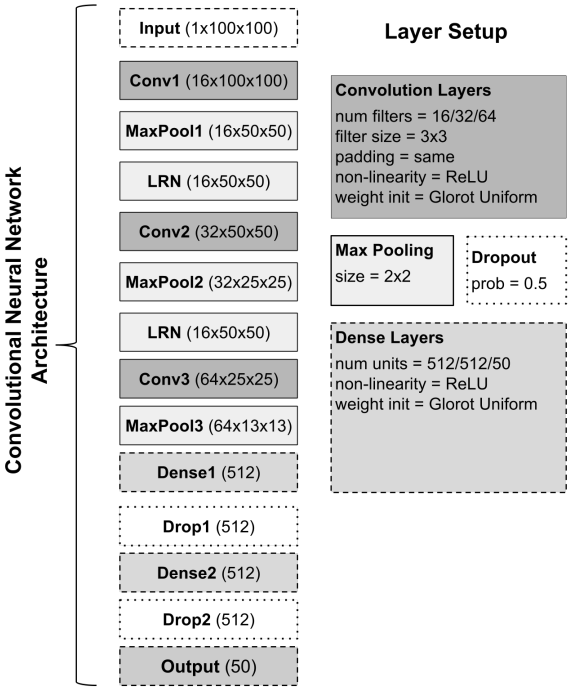


Figure 2. Overview of the Convolutional Neural Network architecture based on Levis [23].

The test phase uses a voting strategy to predict the class of the music score. The CNN predicts a class for each patch of the original image, which was previous extracted and selected (see Section 4.1). The final class attributed to the music score image is the one with the most votes. Figure 3 illustrates the process.

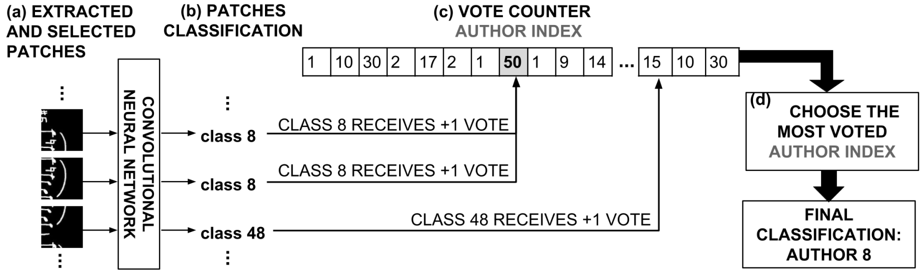

(a) EXTRACTED AND SELECTED PATCHES
(b) PATCHES CLASSIFICATION
(c) VOTE COUNTER
AUTHOR INDEX
1 10 30 2 17 2 1 50 1 9 14 ... 15 10 30
CONVOLUTIONAL NEURAL NETWORK
class 8
class 8
class 48
CLASS 8 RECEIVES +1 VOTE
CLASS 8 RECEIVES +1 VOTE
CLASS 48 RECEIVES +1 VOTE
(d) CHOOSE THE MOST VOTED AUTHOR INDEX
FINAL CLASSIFICATION: AUTHOR 8

Figure 3. Overview of the voting strategy. (a) all patches preprocessed from a specific music score (b) all patches are classified (c) votes are counted (d) the most voted writer wins.

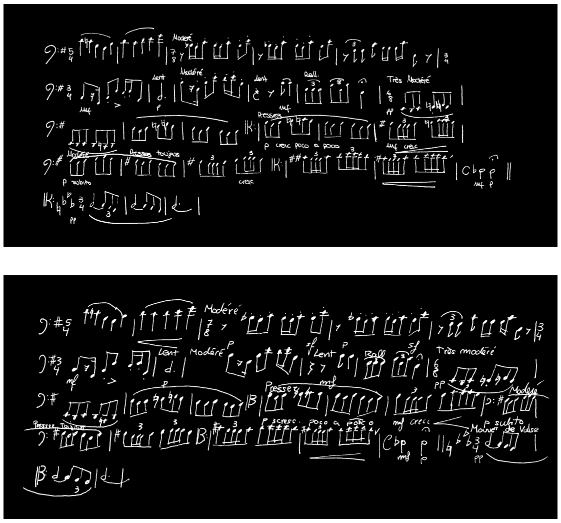


Figure 4. Images extracted from the CVC-MUSCIMA dataset showing samples from two different writers, but from the same music score.

5 Experiments

All experiments done in this work were done in a computer running Ubuntu 14.04 LTS with an Intel Core i7 processor at 3.30GHz and a Nvidia Titan X GPU. The software was developed using the Tensorflow 0.11 framework [1].

The main objective of this work is to evaluate the performance of the Patch-Based CNN using a voting system for writer identification in music scores (see Section 4). In order to evaluate the proposed approach, experiments were done using the CVC-MUSCIMA dataset (see Section 5.1).

5.1 Music Score Dataset

In our experiment, the CVC-MUSCIMA[2] is used. This dataset is a handwritten music score image dataset designed for staff removal and writer identification. Samples from the dataset are shown in Figure 4. In the classification task, there are a total of 50 classes, i.e. authors. Each author transcribes the same 20 music scores. In total, the dataset contains 1.000 images of roughly 3.500×1600 pixels with some variations. The dataset offers images with and without staff lines. In this work, we use only the staff-less images. By using our patch extraction strategy, each music score produces an average of 250 patches.

The CVC-MUSCIMA dataset has predefined train and test folds for cross-validation. However, cross-validation is computationally expensive when using

[1] Available in: `https://www.tensorflow.org/`

[2] Available in: `http://www.cvc.uab.es/cvcmuscima/index_database.html`

Deep Learning methods such as the CNN. Hence, we employ the Hold-Out method. The data is split in the following manner: the first 17 music scores of each writer are used for training, whilst the remaining 3 are used for testing. Hence, 850 images of music scores (227.560 patches) were used to train and 150 images of music scores (45.063 patches) were used to test our approach.

6 Results and Discussion

Since the number of music scores for each writer is the same, the performance metric used to evaluate our approach is the overall accuracy ($A_{CC} = C/N$), which is calculated by dividing the number of correct classifications C by the number of samples N. We also consider the top-1, top-3 and top-5 error rates obtained by our approach.

Figure 5 shows the learning behavior of our model. We can observe that the method can decrease the loss of the train and the test, without the network overfitting. Our approach achieved an loss equal to 2.19 and 2.46, in the train and test. In the Table 1 we show the accuracy of the Patch-Based CNN. The Top-1 accuracy achieved 84%, whilst the Top-3 and Top-5 achieved 96% and 98%, respectively.

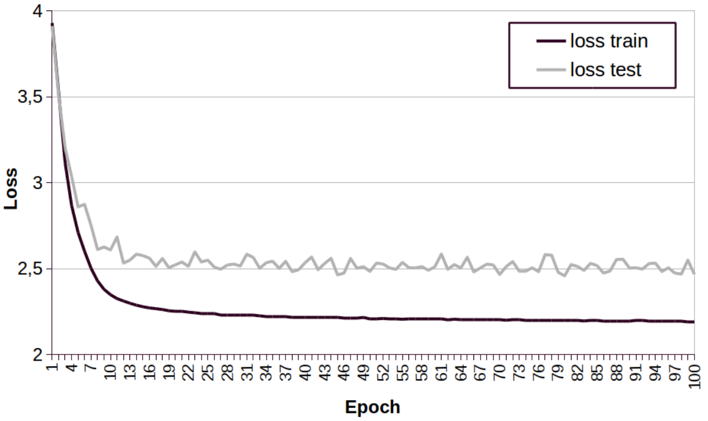
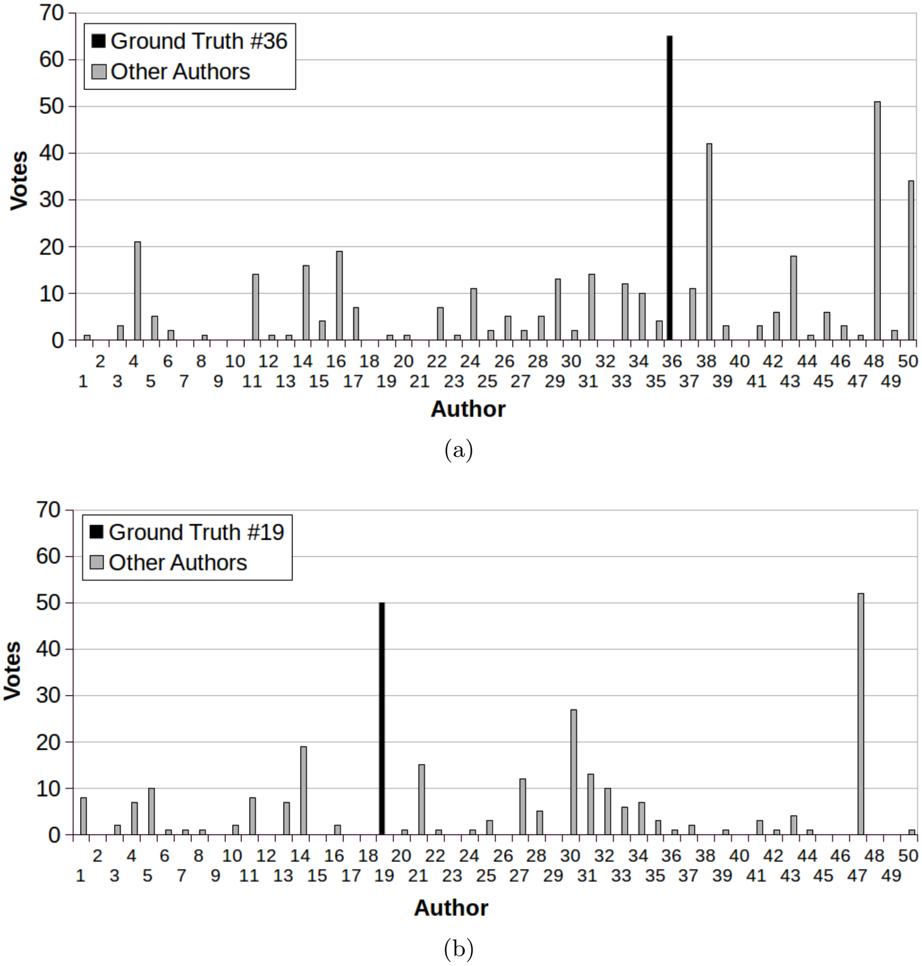


Figure 5. Plot of the loss model on Train and Test sets in each epoch.

Figure 6(a) and 6(b) show the results of the voting system in a correct classification and an incorrect classification, respectively. In the correct classification

Table 1. Top accuracy obtained by the Patch-Based CNN.

	Top-1	**Top-3**	**Top-5**
Accuracy	84%	96%	98%



(a)

(b)

Figure 6. Examples of two vote vectors: (a) a correct classification (b) an incorrect classification.

case, it is possible to observe that the correct author (65 votes) received a significantly higher amount of votes than the other authors. The second most voted author received 51 votes, indicating that there may be similarities in the writing style of the two authors. This indicates that the method is able to capture the peculiarity of the handwriting of the authors. In the wrong case, the method failed to predict the correct author by voting in the incorrect author 52 times. However, the histogram shows that the voting system got close to the correct result, since the correct author received 50 votes. This endorses the results presented in Table1.

7 Conclusion

The Patch-Based CNN approach presented in this work represents an important contribution for solving Writer Classification Problems (WCPs) regarding music scores, since it presents the first implementation of a voting system approach in conjunction with a CNN applied to Music Score Classification.

Considering the results obtained in the experiments presented in this work, our approach showed to be able to learn the nuances of the writings of each author. The method allowed to obtain very satisfactory results for the case study: for the top-3 and top-5 accuracies, the performances are close to 100%. On the other hand, the top-1 accuracy also achieved a satisfactory result (84%) considering that the dataset used in this work contains 50 classes. However, we consider that this performance may be improved in future works by using more complex CNN architectures in combination with the application of strategies to reduce overfitting of deep models such as data augmentation with noise on the inputs.

In a broader sense, we believe that the proposed approach presented in this paper is very promising for all research areas related to WCP. In this sense, a possible future work could aim at the use of the method in other WCP applications or datasets. Other CNN architectures may also be studied in combination with the voting strategy in order to improve the classification performance of the top-1 accuracy. Moreover, sequential approaches such as Recurrent Neural Networks and Long-Short Term Memory Networks may also be studied for classifying sequences of patches. Although the case of study is related to WCP, this approach may also be studied to solve different problems that require the learning of local features to obtain a final prediction.

Acknowledgment

N.Aquino thanks the Organization of the American States, the Coimbra Group of Brazilian Universities and the Pan American Health Organization; H.S.Lopes thanks to CNPq for the research grant no. 311778/2016-0. All authors would like to thank both CAPES and CNPq for the scholarships, as well as NVIDIA and Fundação Araucária.

Bibliography

[1] Fornés, A., Dutta, A., Gordo, A., Lladós, J.: CVC-MUSCIMA: a ground truth of handwritten music score images for writer identification and staff removal. International Journal on Document Analysis and Recognition (IJDAR) **15**(3) (2012) 243–251

[2] Shi, B., Bai, X., Yao, C.: Script identification in the wild via discriminative convolutional neural network. Pattern Recognition **52** (2016) 448–458

[3] Zhang, X.Y., Xie, G.S., Liu, C.L., Bengio, Y.: End-to-end online writer identification with recurrent neural network. IEEE Transactions on Human-Machine Systems **47**(2) (2017) 285–292

[4] Hafemann, L.G., Sabourin, R., Oliveira, L.S.: Learning features for offline handwritten signature verification using deep convolutional neural networks. Pattern Recognition **70** (2017) 163–176

[5] Joshi, P., Agarwal, A., Dhavale, A., Suryavanshi, R., Kodolikar, S.: Handwriting analysis for detection of personality traits using machine learning approach. International Journal of Computer Applications **130**(15) (2015)

[6] Rosenblum, S., Dror, G.: Identifying developmental dysgraphia characteristics utilizing handwriting classification methods. IEEE Transactions on Human-Machine Systems **47**(2) (2017) 293–298

[7] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553) (2015) 436–444

[8] Fornes, A., Dutta, A., Gordo, A., Llados, J.: The ICDAR 2011 music scores competition: Staff removal and writer identification. In: 2011 International Conference on Document Analysis and Recognition (ICDAR), IEEE (2011) 1511–1515

[9] Dalitz, C., Droettboom, M., Pranzas, B., Fujinaga, I.: A comparative study of staff removal algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(5) (2008) 753–766

[10] Hannad, Y., Siddiqi, I., El Merabet, Y., El Youssfi El Kettani, M.: Arabic writer identification system using the histogram of oriented gradients (HOG) of handwritten fragments. In: Proceedings of the Mediterranean Conference on Pattern Recognition and Artificial Intelligence, ACM (2016) 98–102

[11] Khan, F.A., Tahir, M.A., Khelifi, F., Bouridane, A., Almotaeryi, R.: Robust off-line text independent writer identification using bagged discrete cosine transform features. Expert Systems with Applications **71** (2017) 404–415

[12] Roy, P.P., Bhunia, A.K., Pal, U.: HMM-based writer identification in music score documents without staff-line removal. Expert Systems with Applications (2017)

[13] Niitsuma, M., Schomaker, L., Van Oosten, J.P., Tomita, Y., Bell, D.: Musicologist-driven writer identification in early music manuscripts. Multimedia Tools and Applications **75**(11) (2016) 6463–6479

[14] Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2016** (2016) 2424–2433

[15] LeCun, Y., Kavukcuoglu, K., Farabet, C.: Convolutional networks and applications in vision. In: IEEE International Symposium on Circuits and Systems, Piscataway, NJ, IEEE Press (2010) 253–256

[16] Szegedy, C., Liu, W., Jia, Y., SermarXivanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, IEEE Computer Society (2015) 1–9

[17] Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Aistats. Volume 15. (2011) 275

[18] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, Red Hook, NY, Curran Associates, Inc. (2012) 1097–1105

[19] Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv:1207.0580 (2012) 1–18

[20] Ng, A.Y.: Feature selection, L1 vs L2 regularization, and rotational invariance. In: Proceedings of the 21 International Conference on Machine Learning, New York, NY, USA, Advances in Neural Information Processing Systems, ACM (2004) 78

[21] Perlin, H.A., Lopes, H.S.: Extracting human attributes using a convolutional neural network approach. Pattern Recognition Letters **68**(2) (2015) 250–259

[22] Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**(1) (2014) 1929–1958

[23] Levi, G., Hassncer, T.: Age and gender classification using convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (June 2015) 34–42

[24] Kingma, D., Ba, J.: Adam: A method for stochastic optimization., arXiv (2014)