This ECCV 2018 workshop paper, provided here by the Computer Vision Foundation, is the author-created version. The content of this paper is identical to the content of the officially published ECCV 2018 LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/eccv

Towards automated multiscale imaging and analysis in TEM: Glomerulus detection by fusion of CNN and LBP maps

Elisabeth Wetzer[1] [0000−0002−0544−8272], Joakim Lindblad[1,4] [0000−0001−7312−8222], Ida-Maria Sintorn[1,3], Kjell Hultenby[2], and Nataša Sladoje[1,4] [0000−0002−6041−6310]

[1] Uppsala University, Sweden
{elisabeth.wetzer, joakim.lindblad, ida.sintorn}@it.uu.se
[2] Karolinska Institute, Sweden
kjell.hultenby@ki.se
[3] Vironova AB, Stockholm, Sweden
[4] Mathematical Institute of Serbian Academy of Sciences and Arts, Belgrade, Serbia

Abstract. Glomerulal structures in kidney tissue have to be analysed at a nanometer scale for several medical diagnoses. They are therefore commonly imaged using Transmission Electron Microscopy. The high resolution produces large amounts of data and requires long acquisition time, which makes automated imaging and glomerulus detection a desired option. This paper presents a deep learning approach for Glomerulus detection, using two architectures, VGG16 (with batch normalization) and ResNet50. To enhance the performance over training based only on intensity images, multiple approaches to fuse the input with texture information encoded in local binary patterns of different scales have been evaluated. The results show a consistent improvement in Glomerulus detection when fusing texture-based trained networks with intensity-based ones at a late classification stage.

Keywords: Glomerulus detection; Transmission Electron Microscopy; Convolutional Neural Networks; Local Binary Patterns; Digital pathology

1 Introduction

The glomerulus is a structure in the kidney which acts as a filtration barrier for metabolic waste from the bloodstream. A number of diseases, such as minimal change disease, systemic lupus and many others, can affect the glomerulus and have serious impact on the kidneys and their function. Analysis of the thickness of the glomerular basement membrane (GBM), deposits of amyloid fibres, protein or virus-like deposits in the membrane, and foot process effacement, are some of the necessary nephropathological diagnostic procedures. Diagnostically relevant glomerular structures, such as protein fibres and deposits, are of nanometer-scale dimensions, which makes Transmission Electron Microscopy (TEM) the preferred imaging technique for glomerular analysis.

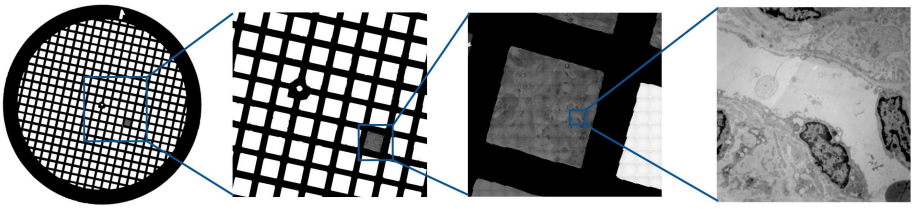


Fig. 1. A series of magnifications illustrating the vast search space for glomeruli on a mesh grid used for TEM with a diameter of 3 mm (left) to a single acquired image of size 16×16 µm (right). For ground truth annotation, to reliably identify a glomerulus, the foot processes along the GBM were used for recognition

The first step in the analysis is detection of glomeruli in the sample. This is typically done at low magnification which allows a relatively large field of view (FOV) of the sample. A pathologist then continues the analysis of zoomed-in regions of interest, i.e., at higher magnification, focusing on the relevant structures of the glomerulus. This multi-scale analysis is currently performed manually, requiring that a pathologist spends a long time at the electron microscope. Figure 1 shows a typical sample mesh used in TEM, and a series of visualizations at increasing magnifications to the level at which the foot processes along the GBM are visible.

Fully automated image acquisition of the relevant portions of the sample (glomeruli) in high magnification would significantly reduce the consumption of precious expert time and allow the expert to perform the diagnosis at his/her choice of time and place, rather than at the microscope.

We suggest a two-step approach to automate the imaging process: (1) Scanning is first performed at low magnification to identify regions of interest, followed by (2) imaging of only those detected regions of interest at high magnification. In this study, we focus on the first step: glomerulus detection in low magnification TEM images. The detection is based on classification of whole (low-mag) images as either part of a glomerulus or as other kidney tissue. An example of such a low-mag image is shown in Fig. 1 (right).

In bright-field microscopy images, glomeruli are recognized and detected by their characteristic texture; a variant of the LBP texture descriptor, named multi-radial color LBP (mcLBP) has recently shown to perform very well on the task [18]. In this study, we combine the ability of LBPs to describe the fine textural details with the classification power of Convolutional Neural Networks (CNNs). We are particularly inspired by the approach proposed in [1,8], where authors compute dense LBP feature maps which, in combination with raw image data, are classified using deep CNNs.

2 Background and preliminaries

2.1 Previous work

The descriptive power of recently suggested mcLBP for glomeruli detection in bright-field microscopy images, [18], results from concatenated histograms of LBPs with different radii, computed for each RGB color channel separately. To further boost the performance and decrease the number of false positive detections, authors train a deep CNN, GoogleLeNet. They observed that the deep learning based approach solely performs worse than the one based on mcLBP. A number of other papers have demonstrated that it is often beneficial to combine hand crafted and learned features [12, 13, 16, 17]. This can be seen as a variation of transfer learning, where the network is helped by additional views of the imaged data.

Approaches to combine the power of machine learning and LBPs in texture-based classification include extraction of histograms of LBP responses over sliding windows (in histopathological whole-slide images), followed by support vector machine (SVM) classification [18]. LBP histograms have also been used in combination with learned features of CNNs [12]. LBP-like features can also be learned, as in [11]. Instead of computing LBP histograms, LBPs can be used as a dense feature extractor, and combined with CNN [1, 8]. Furthermore, LBPs can be interpreted as convolutional layers, with learned parameters [6, 9].

We are following the very promising approach proposed in [1, 8], where authors use dense LBP maps, in combination with the raw image data, as input for a CNN. The generation of LBP codes of an image results in an unordered set of binary codes where the distance between code values of two patterns does not reflect the distance between the patterns. This makes the direct usage of LBP codes unsuitable for CNNs; the discrete convolution operation computes a weighted sum of input values, similar to interpolation, but interpolation of the LBP codes does not have a meaningful interpretation. Therefore, a dissimilarity measure, defined in [8] for all possible codes, is used in multidimensional scaling (MDS), which is then applied to map the unordered set of codes into a metric space. This enables the (meaningful) use of convolutions on LBP maps.

2.2 Preliminaries

Local Binary Patterns (LBP) [14] are among the most successful texture descriptors in image analysis. Over time, a number of variants have been proposed, [10], finding numerous applications, [15].

In general, the LBP code $LBP_{r,p}(c)$, for a pixel c with intensity value g_c, is

$$LBP_{r,p}(c) = \sum_{i=0}^{p-1} s(g_i - g_c)\, 2^i \qquad s(x) = \begin{cases} 1 & x \geq 0, \\ 0 & x < 0 \end{cases} \tag{1}$$

where $g_i \in \{0 \dots p-1\}$ are pixels sampled equidistantly in a circle of radius r in the neighborhood of g_c.

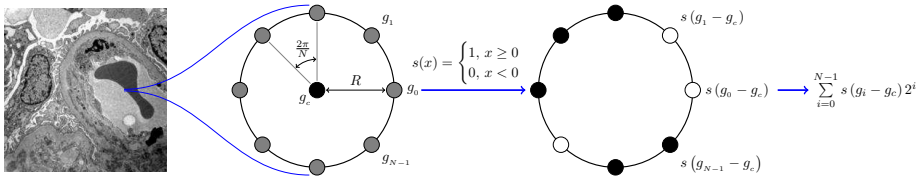


Fig. 2. LBP code sampling for intensity images

LBP codes are most often binned over the full image into a histogram, thereby providing a feature vector of size 2^p for the entire image. In [8] however, an LBP code is generated for every pixel without any following binning. Hence, a value in the range $[0, 2^{p-1}]$ is obtained for every pixel. This texture representation is, however, not well suited as an input to a deep convolutional network. CNNs are based on discrete convolutions which can be seen as a weighted average of their input. On the other hand, LBP codes are binary encoded such that codes with similar numeric values may represent not at all similar patterns. Performing numerical operations such as averaging is therefore not reasonable on a set of LBP codes; they need to be mapped into a metric space first. Such an approach, using multidimensional scaling to map LBP codes into Euclidean space, is proposed in [8].

Multidimensional Scaling (MDS) is a common technique in data science. Using (dis-)similarities of data points, MDS can be used to map the data from an unordered set $\mathbf{X} \subseteq \{2_2^8\}$ into a metric space by numerical optimization [2].

Non-metric multidimensional scaling is performed on the dissimilarity matrix $\Delta = (\delta_{ij}) \in \mathbb{R}_+^{n \times n}$. The so-called representation function $f(\delta_{ij})$ specifies the relation between the dissimilarities and their corresponding metric values $\mathbf{D} = (d_{ij}) \in \mathbb{R}_+^{n \times n}$ which lie in an Euclidean space and approximate a monotonic transformation of δ_{ij}. The resulting optimization problem aims to minimize an objective function referred to as stress. We follow [8] and use non-metric stress normalized by the sum of squares of the inter-point distances, also known as Kruskal's normalized stress-1 criterion, [7]:

$$\text{Stress-1} = \sqrt{\frac{\sum [f(\delta_{ij}) - d_{ij}(\mathbf{X})]^2}{\sum d_{ij}^2(\mathbf{X})}} \tag{2}$$

Dissimilarity Measure To apply MDS to the set of LBP codes, we use one of the dissimilarity measures between the codes suggested in [8]:

$$\delta_{ij} = \delta(P_i, P_j) = \min \left\{ \tilde{\delta}(P_i^0, P_j^0), \tilde{\delta}(\text{rev}(P_i^0), P_j^0), \tilde{\delta}(P_i^0, \text{rev}(P_j^0)) \right\}. \tag{3}$$

Here, $\tilde{\delta}(P_i, P_j) = \|CDF(P_i) - CDF(P_j)\|_1$, where $CDF(P)$ is the cumulative distribution function of bit values; this approximates the Earth Mover's Distance

between the strings (more details about the efficient computation can be found in [8]). P^0 is the concatenation of the binary string P and an additional bit of 0 and rev(P) the rearrangement of a string P in reverse order.

LBP maps are obtained by MDS (using Eq.(2)) applied to LBP codes computed for every image pixel. We follow recommendations from [1,8] and map the LBP codes into a 3-dimensional space. An example of a resulting LBP map is shown in Fig. 3, visualized as an RGB image.

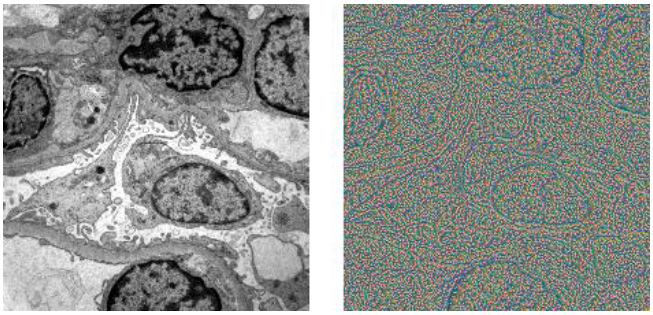


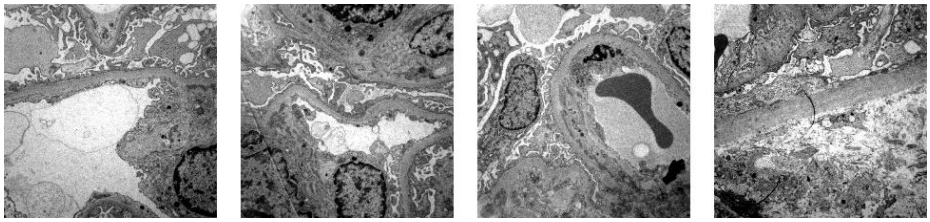
Fig. 3. Intensity image of Glomerulus and its corresponding LBP map in 3D metric space

3 Method

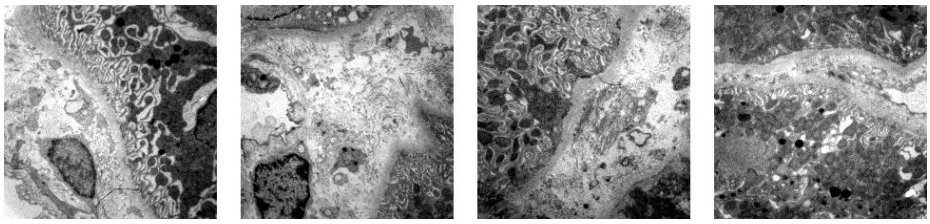
3.1 Dataset

16-bit intensity images of the size 2048 $\times$ 2048 pixels were acquired using MiniTEM[5], a desktop, low-voltage (25keV) transmission electron microscope. The dataset consists of 494 images, grouped in two sets used for training and testing. The sets were independently acquired at different occasions using the built-in automatic imaging function in MiniTEM. The *training set* consists of 260 images, 70 of which have been marked as containing glomerulus specific structures, and 190 to contain other kidney tissue. The field of view (FOV) covered by one image is 16 µm, yielding a pixel size of 7.8 nm. The *test set* consists of 56 images containing glomerulus tissue, and 178 images of other kidney tissue (from two different samples). Example images of both classes, i.e., glomerulus and non-glomerulus, are shown in Fig. 4, illustrating the difficult task of distinguishing the two. Ground truth annotation was done on an image level based on the visual detection of foot processes.

[5] Vironova AB, Stockholm



(a) Glomerulus tissue



(b) Other tissue structures

Fig. 4. Example images from the two classes

3.2 Architecture

We compare two architectures for the CNN models: VGG16 and ResNet50. They are trained from scratch on either the raw image data or the LBP maps. We evaluate fusion of raw and LBP data at three different depths of the networks.

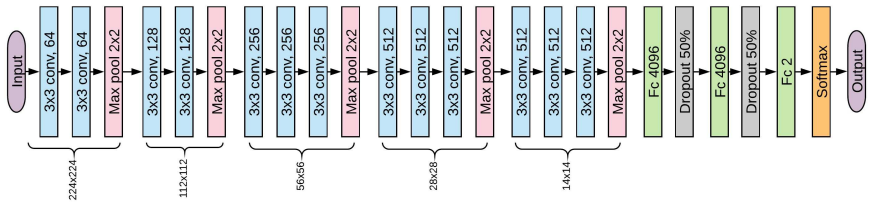
Data augmentation, consisting of $0^\circ, 90^\circ, 180^\circ, 270^\circ$ rotations and mirroring, is performed for all training data, leading to $8\times$ data augmentation without interpolation. LBP codes are computed on the TEM intensity images. Following the LBP computation, the input data (intensity images as well as LBP maps) were resized to 224×224 pixels using nearest neighbor interpolation.

VGG16-like Architecture: The architecture used in the experiments is a modification of the VGG16 network [19], shown in Fig. 5(a). A batch normalization layer is introduced after each convolutional layer.

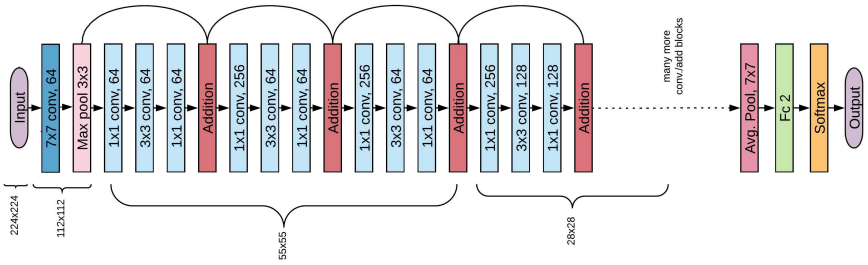
ResNet50: ResNet50 is a residual neural network with a total of 177 layers including batch normalizations and activation layers. The layers are learning residual functions with respect to the layer inputs, [5].

We have investigated three different fusion strategies – early, mid and late fusion – to combine texture and intensity information. To enhance multi-scale descriptive power of LBPs, we observe $LBP^1_{1,8}$, $LBP^2_{2,8}$, and $LBP^3_{3,8}$, and create their corresponding maps.

Early Fusion: In the early fusion model, the raw image layer is stacked with the three layers of (3D) LBP maps and fed into the input layer of the CNN. The



(a) VGG16 – a batch normalization layer was introduced after each convolutional layer



(b) ResNet50 architecture as introduced in [5]

Fig. 5. Used architectures. Early fusion takes place in the input layer, mid fusion in the second fully connected layer (VGG16 only), and late fusion after the softmax layer.

intensity image and LBP maps are subsequently jointly used in training using cross-entropy loss and stochastic gradient descent. In the multi-scale experimental setup, the raw image layer is stacked with the, in total, nine layers of LBP maps corresponding to the varying radii ($r \in \{1, 2, 3\}$) in LBP extraction.

Mid Fusion: The mid fusion model is only tested for the VGG16-like architecture. It uses a two-stream architecture; one CNNs is trained on the normalized intensity images, the other on the (single scale) 3-layer LBP maps. Once the two networks are trained, the outputs of the second fully connected layers of both architectures are concatenated, resulting in 2×4096 features. A linear SVM is then trained on the resulting 8192-feature vectors.

Late Fusion: Two CNNs are independently trained, one using the normalized intensity data as input, the other the 3-layer LBP maps. As in [1], the output probabilities of the softmax layers of the two networks are concatenated and a linear SVM is trained to classify the data based on such 4-feature vectors. For the multi-scale setup, the outputs of four networks are fused, thus resulting in 8-feature vectors for the SVM.

Table 1. Classification accuracies (with std.dev.) for the different approaches.

Architecture	Input	No fusion	Early fusion	Mid fusion	Late fusion
VGG16	Intensity	0.907 (0.061)			
VGG16	$LBP_{1,8}$	0.843 (0.014)	0.839 (0.076)	0.971 (0.013)	0.972 (0.017)
VGG16	$LBP_{2,8}$	0.941 (0.012)	0.770 (0.046)	0.970 (0.011)	0.977 (0.019)
VGG16	$LBP_{3,8}$	0.915 (0.013)	0.759 (0.068)	0.972 (0.009)	0.969 (0.007)
ResNet50	Intensity	0.964 (0.016)			
ResNet50	$LBP_{1,8}$	0.926 (0.018)	0.728 (0.004)		**0.984** (0.003)
ResNet50	$LBP_{2,8}$	0.929 (0.073)	0.724 (0.008)		0.979 (0.007)
ResNet50	$LBP_{3,8}$	0.946 (0.016)	0.731 (0.018)		**0.984** (0.006)
VGG16	Multiscale		0.863 (0.026)		0.983 (0.008)
ResNet	Multiscale		0.857 (0.021)		0.980 (0.004)
VGG16 Transfer	Intensity	0.877 (0.062)			
ResNet50 Transfer	Intensity	0.963 (0.005)			
VGG16 ResNet50 Ensemble	Intensity				0.970 (0.005)
SVM, baseline	$LBP_{2,8}$	0.752			

4 Evaluation

All models are trained from scratch for 20 epochs using stochastic gradient descent with momentum of 0.9, a learning rate of 0.001, an L_2 regularization of 10^{-4} and a mini-batch size of 16. Average accuracy (ratio of the correctly identified test samples and their total number) over seven runs of CNN experiments for two types of architectures, VGG16 and ResNet50, are reported in Tab. 1 and Fig. 6. We present results obtained by networks trained solely on one type of input (intensity images or LBP maps), as well as results obtained by different methods of fusion (early, mid, or late) of the intensity images and LBP maps, with different (indicated) parameters. *Multiscale* refers to the fusion of the LBP maps of three different radii with the intensity image data (4 networks fused).

For comparison, transfer learning on the intensity data is evaluated for the VGG16, as well as the ResNet50 architecture. Both networks were pretrained on ImageNet [3], whereafter the last fully connected layer (in each) was retrained on the glomerulus data. Results are included in Table 1. The transfer learning performance is slightly lower than the from-scratch performance; we assume this is due to TEM images differing significantly from ImageNet data.

The effect of architecture ensembles has been investigated for reference by training a linear SVM on the softmax layer output of the VGG16 like architecture and ResNet50 architecture which were trained from scratch on the intensity images. The approach is similar to the late fusion, but with two different architectures on the same input, instead of the same architecture with two different input sources. It improves the outcome slightly compared to a single

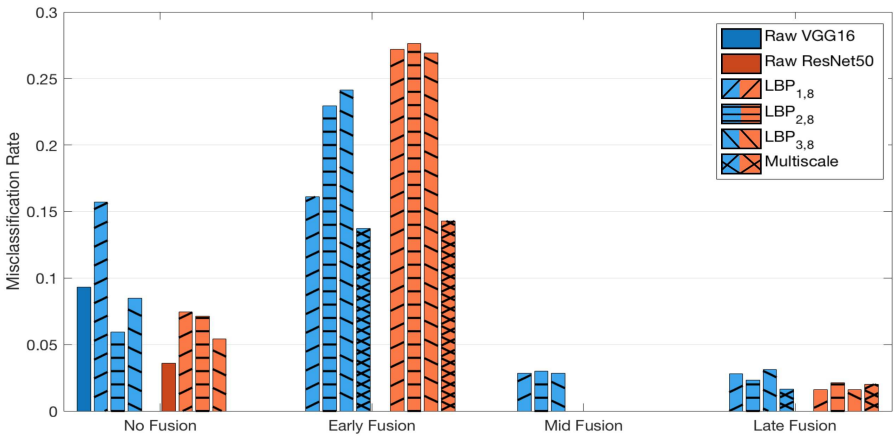
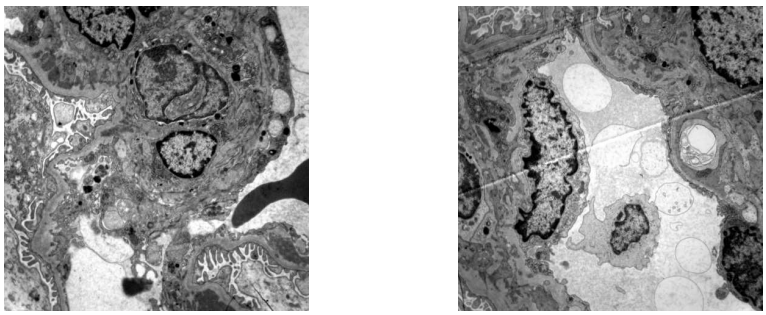


Fig. 6. Misclassification rate of different approaches, VGG16 in blue, ResNet50 in orange. Late fusion models, in particular based on ResNet50, give best results

architecture performance, but does not reach the performance of the multiple input ensembles (fusion) utilizing the LBP maps.

As a reference performance, classification based on (multiple versions of classic) LBP histograms, using a linear SVM classifier, is performed. We observe *uniform* $LBP_{1,8}$ and $LBP_{2,8}$, as well as two *rotation-invariant* versions of $LBP_{1,8}$ and $LBP_{2,8}$: by bit-wise shifting [14], and by using the discrete Fourier transform [4]. The best accuracy, reached by the DFT rotation-invariant $LBP_{2,8}$, is 0.752, which is considerably worse than the performance of the proposed method.

During manual post-validation of the results, one image, consistently classified as a false positive Glomerulus detection (and included in the quantitative



(a) Glomerulus missed (b) Detected, mislabeled Glomerulus

Fig. 7. The only two misclassifications of the ResNet50 Late Fusion of intensity images and $LBP_{3,8}$: (a) False negative and (b) False positive detection (confirmed as mislabeled)

evaluation as such), is found to have been incorrectly labelled during ground truth annotation. The only two misclassified images are shown in Fig. 7.

5 Conclusions

Our results show a consistent improvement in classification accuracy when texture information in form of LBP maps is fused with intensity information and classified using CNNs, compared to only relying on either classic SVM classification of LBPs, or CNN classification based on the intensity information only. Mid and late fusion exhibit similar performance for the VGG16 architecture, and both yield clearly better results than the early fusion strategy. ResNet50 exhibits superior performance to VGG16 when applied to a single type of input, in all cases but one (for $LBP_{2,8}$). The early fusion of input sources for ResNet50 performs the worst for all LBP maps, while the late fusion gives very good results, of which all exceed the accuracy achieved by VGG16. The multiscale approach of fusing LBP maps of varying LBP radii and intensity information gave the best results among all experiments for the early fusion setup, yet the late fusion yields overall best results. For the late fusion we cannot draw any clear conclusion about optimal LBP radius, the different radii perform roughly equally well. The multi-scale approach has higher impact on VGG16. We confirm that delaying the fusion and reduction of features to the very end leads to the best results for this application. Our promising preliminary results encourage continuation of the study on a larger dataset.

Acknowledgment

This work is supported by VINNOVA, MedTech4Health grants 2016-02329 and 2017-02447, the Ministry of Education, Science, and Techn. Development of the Rep. of Serbia (proj. ON174008 and III44006), and the Centre for Interdisciplinary Mathematics, Uppsala University.

References

1. Anwer, R.M., Khan, F.S., van de Weijer, J., Molinier, M., Laaksonen, J.: Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. ISPRS Journal of Photogrammetry and Remote Sensing **138**, 74–85 (2018)
2. Borg, I., Groenen, P.J.F.: Modern Multidimensional Scaling – Theory and Applications. Springer New York (2005)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conf. Comp. Vis. Patt. Rec. pp. 248–255 (2009)
4. Fernández, A., Ghita, O., González, E., Bianconi, F., Whelan, P.F.: Evaluation of robustness against rotation of LBP, CCR and ILBP features in granite texture classification. Machine vision and Applications **22**(6), 913–926 (2011)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)

6. Juefei-Xu, F., Boddeti, V.N., Savvides, M.: Local binary convolutional neural networks. In: Comp. Vis. Pat. Rec.(CVPR), IEEE Conf. on. vol. 1, pp. 19–28. IEEE (2017)
7. Kruskal, J.B., Wish, M.: Multidimensional scaling. Sage University Papers Series. Quantitative Applications in the Social Sciences **11**, 234–778 (1978)
8. Levi, G., Hassner, T.: Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: Proc. of ACM Int. Conf. on Multimodal Interaction. pp. 503–510. ACM (2015)
9. Li, L., Feng, X., Xia, Z., Jiang, X., Hadid, A.: Face spoofing detection with local binary pattern network. Journal of Visual Communication and Image Representation **54**, 182–192 (2018)
10. Liu, L., Fieguth, P., Guo, Y., Wang, X., Pietikainen, M.: Local binary features for texture classification: Taxonomy and experimental study. Patt. Rec. **62**, 135 – 160 (2017)
11. Lu, J., Liong, V.E., Zhou, X., Zhou, J.: Learning compact binary face descriptor for face recognition. IEEE Trans. on PAMI **37**(10), 2041–2056 (2015)
12. Majtner, T., Yildirim-Yayilgan, S., Hardeberg, J.Y.: Combining deep learning and hand-crafted features for skin lesion classification. In: Image Processing Theory Tools and Applications (IPTA), Int. Conf. on. pp. 1–6. IEEE (2016)
13. Nahid, A.A., Kong, Y.: Histopathological breast-image classification using local and frequency domains by convolutional neural network. Information **9**(1), 19 (2018)
14. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. on PAMI **24**(7), 971–987 (2002)
15. Pietikainen, M., Hadid, A., Zhao, G., Ahonen, T.: Computer vision using local binary patterns. Springer London, UK (2011)
16. Rezaeilouyeh, H., Mollahosseini, A., Mahoor, M.H.: Microscopic medical image classification framework via deep learning and shearlet transform. Journal of Medical Imaging **3**(4), 044501 (2016)
17. Sadanandan, S.K., Ranefall, P., Wählby, C.: Feature augmented deep neural networks for segmentation of cells. In: Proc. European Conf. on Comp. Vision. pp. 231–243. Springer (2016)
18. Simon, O., Yacoub, R., Jain, S., Tomaszewski, J.E., Sarder, P.: Multi-radial LBP features as a tool for rapid glomerular detection and assessment in whole slide histopathology images. Scientific reports **8**(1), 2032 (2018)
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)