Learning to Recognize Irregular Features on Leather Surfaces

by

Masood Aslam,[1] Tariq M. Khan,[2*] Syed Saud Naqvi,[1] Geoff Holmes[3] and Rafea Naffa[3]
[1]*Department of Electrical and Computer Engineering, COMSATS University Islamabad, Islamabad Campus*, Islamabad 45550, Pakistan
[2]*School of Information Technology, Deakin University*, Geelong, VIC 3217, Australia
[3]*NZ Leather and Shoe Research Association (LASRA),* Palmerston North 4414, New Zealand

Abstract

As part of industrial quality control in the leather industry, it is important to identify the abnormal features in wet-blue leather samples. Manual inspection of leather samples is the current norm in industrial settings. To comply with the current industrial standards that advocate large-scale automation, visual inspection based leather processing is imperative. Visual inspection of irregular surfaces is a challenging problem as the characteristics of the abnormalities can take a variety of shape and color variations. The aim of this work is to automatically categorize leather images into normal or abnormal by visual analysis of the surfaces. To achieve this aim, a deep learning based approach is devised that learns to recognize regular and irregular leather surfaces and categorize leather images on its basis. To this end, we propose an ensemble of multiple convolutional neural networks for classifying leather images. The proposed ensemble network exhibited competitive performance obtaining 92.68% test accuracy on our own curated leather images dataset.

Introduction

In the production of high quality wet-blue leather, automation of defect inspection and classification is highly desirable. Inspection is carried out at different stages of production. It is clear that more leather abnormalities can be prevented by earlier identification. To date, the industry's inspection of abnormal leather surfaces relies primarily on human vision. However, because of human error, manual inspection is strongly subjective, which could eventually lead to low productivity and an undesired false detection rate. With the developments of artificial intelligence (AI) in recent years, its practical applications in visual-based detection in many industrial sectors have been successfully demonstrated.[1-3] Also, with the rise of industry standard 4.0,[4] AI based visual inspection of leather samples is imperative.

Visual inspection based leather categorization is a challenging problem as the appearance of the irregularities is highly varying in nature. Hides and skin defects are usually known as antemortem (before animal death), post-mortem (fault after animal death) and as defects of processing. Brand marks, pox marks, tick marks, insect bites, bruises, growth marks, scratches, etc. are potential defects that arise during the animals' lifespan. These defects are distinguished by a range of shapes, colors, and textures. The data shift problem on the test set and similar visual appearance of normal and irregular leather surfaces further make the problem multifaceted.

AI deals with developing theory, methods and systems that can enable machines to mimic human intelligence. Neural networks are a small step in this direction that attempt to mimic the human brain and its functionality. The area of artificial intelligence which deals with imitating the behaviour of human vision is image processing. While other AI techniques have also played their roles, neural networks have come a long way in replicating simple human vision tasks including handwritten digit recognition, number plate reading and visual recognition.[5] Neural networks have demonstrated their efficacy on a wide range of image processing applications including image enhancement, compression, image segmentation, object recognition and image understanding.[5]

In the past non neural network based AI methods have been investigated for visual inspection of leather.[6,7] While they achieve considerable performance, the current state of methods is far from reaching a generic solution that can meet the needs of industrial scale visual inspection. Moreover, convolutional neural networks (CNN) based methods (that are a special type of neural networks) have not been explored to their fullest potential for visual inspection of leather. A valid reason for this is the lack of data sets which is a major impediment to progress in this area. The previous studies do not make their data available for comparative evaluation.

The aim of this paper is to design a system capable of automatic classification of leather images as normal or flawed. The proposed system can be easily adapted to automated leather hide categorization by leveraging its robust image-by-image classification capability. Such a system can act as a support system for the experts and aid in bias free, rapid categorization of leather samples. The major objectives to achieve the aim of the paper are as follows:

Corresponding author email: tariq_mehmood@comsats.edu.pk
Manuscript received November 27, 2020, accepted for publication February 4, 2021.

- systematic ensembling of state-of-the-art CNNs and their adaptation for leather image categorization,
- comparative evaluation of the proposed method with previous stat-of-the-art machine and deep learning based approaches in terms of widely accepted classification performance metrics,
- introduce a high resolution, wet-blue leather image dataset for benchmark comparative evaluation of methods.

In this work, we propose an ensemble convolutional neural network that is designed systemically thorugh empirical evaluations for robust leather sample categorization. We also introduce a new high-resolution wet-blue leather image dataset consisting of normal and defective leather samples. The images are acquired in a controlled environment with a digital camera device. The major challenges in image acquisition using a digital camera device include proper illumination conditions, proper distance from the leather surface and managing the field of view and ensuring high resolution of images. Other important factors are stability of the acquisition device to exclude the possibility of image artefacts. The dataset introduced in this work was curated by taking into account all the above-mentioned issues.

The major contributions of the proposed work are:

- a new ensemble method for robust leather sample classification,
- a thorough comparative evaluation of the proposed method with nine benchmark machine and deep learning based methods,
- introduction of a new high-resolution leather images dataset for stimulating research in the field.

As explained earlier, wet-blue leather has multiple types of defects, however, in this work we are only interested in classifying leather images as normal or abnormal. In this paper, we will use the terms defects and irregular features interchangeably for abnormal leather surface regions (i.e., cuts in our case). The visual appearance of these abnormal image regions is characterized by a wide variety of shapes, textures, scales, spatial locations, and color variations.

The rest of the paper is organized as follows. Section 2 contains a literature review on leather defect classification. Section 3 describes the proposed method. Section 4 explains experimental design. In section 5 all results are presented. Class Activation Maps are explained in section 6. Finally, we conclude our work in section 7.

Literature Review

In this section, we review several machine learning based methods proposed in the literature for leather image classification. To identify abnormal features, Chishti et al. proposed LM-trained multi-layer perceptron neural network structure optimization algorithms have been developed. The suggested results of the method have better accuracy than existing LM-based classifiers without neural structure optimization. Dataset images for wet blue leather and rawhides of the 11 most common features are provided with the classifier results. The algorithm classifies wet blue leather defects with 98.73% accuracy, 97.85% precision, and 94.14% sensitivity.[6] Deng et al. proposed a method to classify surface abnormalities on the whole piece of the leather automatically and objectively, based on a parameter optimized residual network is proposed. They used ResNet-50 and optimized two of the network parameters, the size of the data set and the size of the sliding patch window, are optimized. The size of the data set is obtained by achieving the tradeoffs between the evaluated workload and the classification accuracy. The classification accuracy of the applied reaches 94.6%.[7] Recently, Aslam et al. suggested a method to classify good leather and defected leather images. For the classification task, an ensemble architecture EfficientNet-B3+ResNext-101 is used. The proposed algorithm was able to achieve an AUC of 81.9%.[8]

The Proposed Method

In this work, we investigated deep learning architectures for classification of abnormal surface features in wet-blue leather. Images were acquired using a Nikon Coolpix P300 camera. Training images were employed to augment new image variations to improve the generalization of classifiers on unseen samples. The classifiers were trained using augmented data and train and validation accuracies were computed. The best weights of the trained models were stored for the inference stage. In the inference stage, the test data and the chosen trained model was employed to compute the model predictions, which were then utilized for performance evaluation of the model at the test stage.

Data Augmentation

In order to train the network with different variations of the input images by artificially generating new images for the training, the data augmentation module was added to our workflow. In order to minimise network overfitting and enhance model generalisation, the data augmentation effect has been practically demonstrated.

Horizontal flipping, vertical flipping, rotation using a random angle in the range 30 to 150 and a random zoom factor in the range 0.1 to 0.8 were the chosen augmentation methods in our experiments. With a probability of 0.5, all the transformations were applied. After the data augmentation process, a total of 1557 images were collected. The training data was split randomly into training, validation and test sets with a 60:20:20 ratio. Consequently, for the training collection, 1040 images were used, while the remainder were split into the validation and test sets. The number of defective and non-defected pictures of leather were also held equal, rendering the issue of classification a balanced one. Apart from data augmentation strategies, no particular pre-processing function was applied to the original images.

Convolutional Neural Network Ensembles

Ensemble approaches combine several classifiers, and it has been found that it is possible to obtain greater precision results than a single classifier. For an ensemble, well-known approaches include boosting, bagging and stacking. Stacking combines the outputs of a number of base learners and allows another algorithm , called the meta-learner, to make the final predictions. A super-learner is another technique which calculates the final predictions by finding the optimal weights of the base learners by minimising a loss function based on the cross-validated performance of the learners. Majority voting is an ensemble method that counts all the predicted labels of the base learners, and records the label with the highest number of votes as the final prediction. Another strategy is to measure the optimum weights of individual simple learners. Average voting that produces labels afterwards by calculating the average probabilities of the softmax class or predicted labels for all the base learners is the most common ensemble technique used in neural networks.

In this work, we experimented with various combinations of standard state-of-the-art networks, including VGG-16, ResNet-50, Inception-V3 and Inception-ResNet-V2, in order to find the ideal collection for classifying leather features. The average voting-based ensemble of Inception-V3+ResNet50 was selected as the proposed architecture in this study due to its superior performance and confidence in predictions. The architectural level diagram of the proposed ensemble network is shown in Figure 1.

Ensemble techniques have proven in previous works to be the tool of choice in both related and unrelated image domains. Ensemble approaches combine several classifiers, and it has been found that it is possible to obtain greater precision results than a single classifier. We put together state-of-the-art designs such as Inception-V3, VGG-16, ResNet-50 and Inception-ResNet-V2. Since no representative ensemble methods are discussed for leather image classification in literature, ensemble combinations of two network architectures are exhausted and combinations are chosen in this work that stand out in terms of learned representations. For related domain classification tasks, similar ensemble networks have demonstrated state-of-the-art efficiency. Neural networks are nonlinear and have a high variance, ensemble learning combines the predictions from various architectures to reduce variance of prediction.

Figure 1, consists of two different convolutional neural network architectures, i.e. ResNet-50 and Inception-V3. Input image is fed separately to both the models. Both networks have a Global Average Pooling (GAP) layer as their outputs (for details about GAP, please refer to Section 3.4). In order to calculate the final prediction, the outputs of the networks are fed into the probabilistic averaging layer. The output of the probabilistic averaging layer is passed through a softmax layer that categorizes the input image as normal or defective.

Setting Up the CNN and Training Process

Our model was implemented using Keras deep learning framework 2.1.4. Stochastic Gradient Descent (SGD) and the ADAM optimizers were investigated.[9,10] The momentum rate equal to 0.9 for SGD and set the learning rate to 0.001. We adopted a dynamic learning rate which was divided by 1×10-3 every epoch with an initial value of 1×10-2 for Inception V3 experiments. The training batch size used for Inception-V3+ResNet-50 was 4. We used a binary cross entropy as loss function denoted by (L_{CE} because there are only two classes.[11] The subscript in L_{CE} stands for cross entropy. Our loss function takes the following form

$$L_{CE}(\mathrm{y}, \hat{\mathrm{y}}) = -\frac{1}{N}\sum_{i=0}^{N}(y_i log\,(\hat{y}_i) + (1-y) log(1-\hat{y}_i)) \qquad (1)$$

where y is the label, ŷ denotes the predicted probability and *log* is the natural logarithm.

All images in the dataset were re-sized to 500×375 pixels before training the CNN model to preserve the information in the image and reduce the computational cost of processing. Hence, the input

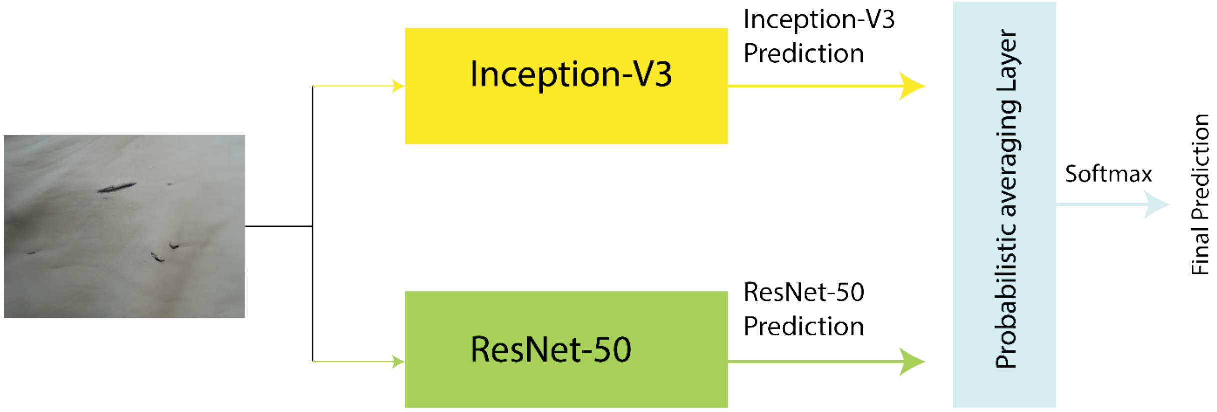

Figure 1. Block level diagram of the proposed ensemble architecture.

Table 1

Comparison of models parameters with and without the GAP layer.

Serial #	Model	Original Parameters	Parameters with GAP
1	Inception-V3	23 Million	21.8 Million
2	Inception-ResNet-V2	55 Million	54 Million
3	ResNet-50	25 Million	23 Million

shape of the CNN is 500×375×3, where three represents the number of RGB channels in the image. We employed the pre-trained weights on ImageNet for the convolutional blocks as an initialization for the network parameters, which resulted in a faster convergence.[12] We also added Global Average Pooling layer and replaced dense layer and flatten layer. The elimination of all these trainable parameters also reduces the tendency of over-fitting, which needs to be managed in fully connected layers by the use of dropout.

Global Average Pooling

One of the most important requirements for visual inspection based leather defect categorization is the large size of the input images. Large image sizes are essential as irregular features can occur at very small scales and equally contribute to the defectiveness of the leather sample. Therefore, the CNN architectures designed for this task must be able to cope with large image sizes. This large image size support is not very common in state-of-the-art CNNs reported in literature for classification tasks. It turns out that large image sizes can only be supported by models that have a relatively lower number of parameters. One such recent state-of-the-art architecture, which has proven to be highly useful at the large scale visual recognition challenge and has an architecture targeted to mobile vision is the Inception-V3 architecture.[13] The Inception-V3 is a well-known state-of-the-art method for multi-class classification with a low computational cost and only consists of 25 million parameters.[13] The ResNet-50 architecture can be carefully optimized to minimize the number of parameters without compromising its established state-of-the-art classification performance in related tasks.

Despite its low computational cost, the proposed Inception-V3+ResNet-50 architecture could not be trained in an end-to-end fashion on leather images greater than 500×375×3 using a multi-gpu hardware resource. To counter this issue, we considered various global pooling layers instead of the fully connected layer to reduce the number of parameters and shed the computational load. Table 1 presents a comparison of model parameters, where the original parameters depict the total parameters (with the fully connected layer), while the parameters with GAP represents the total number of model parameters when the GAP layer is employed. If we look at Inception-V3 parameters, the number of parameters are reduced by about 1 million. Also, in the case of Inception-ResNet-V2, the difference is about 1 million as well. In ResNet-50 the difference is 2 Million parameters, which is slightly more than Inception-V3 and Inception-ResNet-V2.

Four well-known pooling strategies from past works including simpler schemes such as global max pooling,[14] global average pooling,[15] and more complex strategies including log sum exponential (LSE) pooling[16] and max-min pooling[17] were considered. Considering that V^C represents a map from the final convolutional volume of the architecture, we can define all the pooling strategies. The global max pooling strategy denoted by y^cM as defined in,[14] is given as:

$$y^cM = max_{i,j}V^C_{ij}\forall c\in \mathrm{C} \tag{2}$$

Where C represents the number of scores in map and c is the maximum score in that map. The maximum location in the map hypothetically provides the location of the object or the abnormal region in our case.

Similarly, global average pooling (denoted by y^cA), which is another simple pooling scheme, can be defined according to[15] as

$$y^cA = \frac{1}{N}\sum_{i,j}V^C_{ij} \tag{3}$$

Where N is the total number of samples, V^C represents a map from the final convolutional volume of the architecture. It advocates that the location of the region of interest is the global average of the maps instead of the maximum as in global max pooling.

Given a hyperparameter β, the LSE pooling strategy[16] (denoted by y^cLSE) can be expressed as

$$y^cLSE = \frac{1}{\beta}\log(\frac{1}{N}\sum_{i,j}(\exp(\beta V^C_{ij}) \tag{4}$$

where β provides a trade-off between choosing the maximum versus the average values for pooling, *log* is the natural log and *exp* represents the exponential function. In some sense, the LSE pooling strategy provides as trade-off between the global average and max pooling strategies to locate the object of interest. Finally, the min-max pooling strategy based on the *k* highest ($S_{top}(V^c)$) and the *m* lowest ($S_{low}(V^c)$) scoring regions is expressed mathematically according to[17] as

$$y^cMax - Min = S_{top}(V^c) + S_{low}(V^c) \tag{5}$$

$$S_{top}(V^c) = \max h\sum_{i,j}h_{ij}V^C_{ij},\ s.t.\sum_{i,j}h_{ij} = k \tag{6}$$

$$S_{low}(V^c) = \min h\sum_{i,j}h_{ij}V^C_{ij},\ s.t.\sum_{i,j}h_{ij} = m \tag{7}$$

where **h** is a vector responsible for the selection of the candidates. The main idea of the max-min pooling strategy is that multiple regions

Table II

Computational complexity of pooling strategies

Pooling Strategy	Proposed by	Big-O complexity
Global max pooling y^cM	Oquab et al.[14]	$O(n^2)\,t$
Global average pooling y^cA	Zhou et al.[15]	$O(n)$
LSE pooling y^cLSE	Pinheiro et al.[16]	$O(log(O(k(O(n)))))$
Max-min pooling $y^cMax-Min$	Durand et al.[17]	$O(n^{2)}--O(lp)^b$

that hypothesize the location of the object of interest are combined to form the final prediction.

Two important factors that need to be considered when selecting a pooling strategy are the accuracy of the network and the computational overhead of the scheme. In our experiments, all four pooling strategies obtained relatively similar results in terms of accuracy, therefore, we considered computational complexity of the operations to select a particular pooling scheme. Table II, presents the computational complexity of the pooling strategies considered in this work. As the more complex LSE and Max-min pooling schemes have relatively much higher computational complexity with almost similar accuracy as compared with the other two schemes, the simple max and average pooling schemers were considered in this work. Owing to its higher accuracy in our experiments and lower computational complexity according to Table II, global average pooling (GAP) y^cA was considered to be the preferred choice. The overall complexity of the max-min pooling strategy is $O(2n^2)$, however, the parameter h needs optimization, which adds a term $O(lp)$ for each optimization iteration for p number of derivatives.

Experimental Design

Dataset

The original dataset consists of RGB wet-blue leather images with a resolution of 4000×3000 in JPEG format. The images were collected by Nikon Coolpix P300 camera with a 12MP AF sensor. A total of 60 images including 30 normal samples and 30 defected samples were curated to form the original dataset. Equal representation of normal and defective samples was collected to balance the classes. The original images were augmented as explained in Section 3.2 to form a total of 1557 images, out of which, 1040 images were used for training and the rest constituted the validation and test sets.

Figure 2 presents representative examples of normal and defective images along with magnified regions of abnormal and normal leather surfaces. Abnormal or irregular leather surfaces are distinguishable from a normal surface based on the color, texture and shape of the defects characterizing them. For instance, the abnormal surfaces shown with red boxes in Figure 2 contain cuts and white spots, which are highly varying in appearance due to their color, texture and shapes.

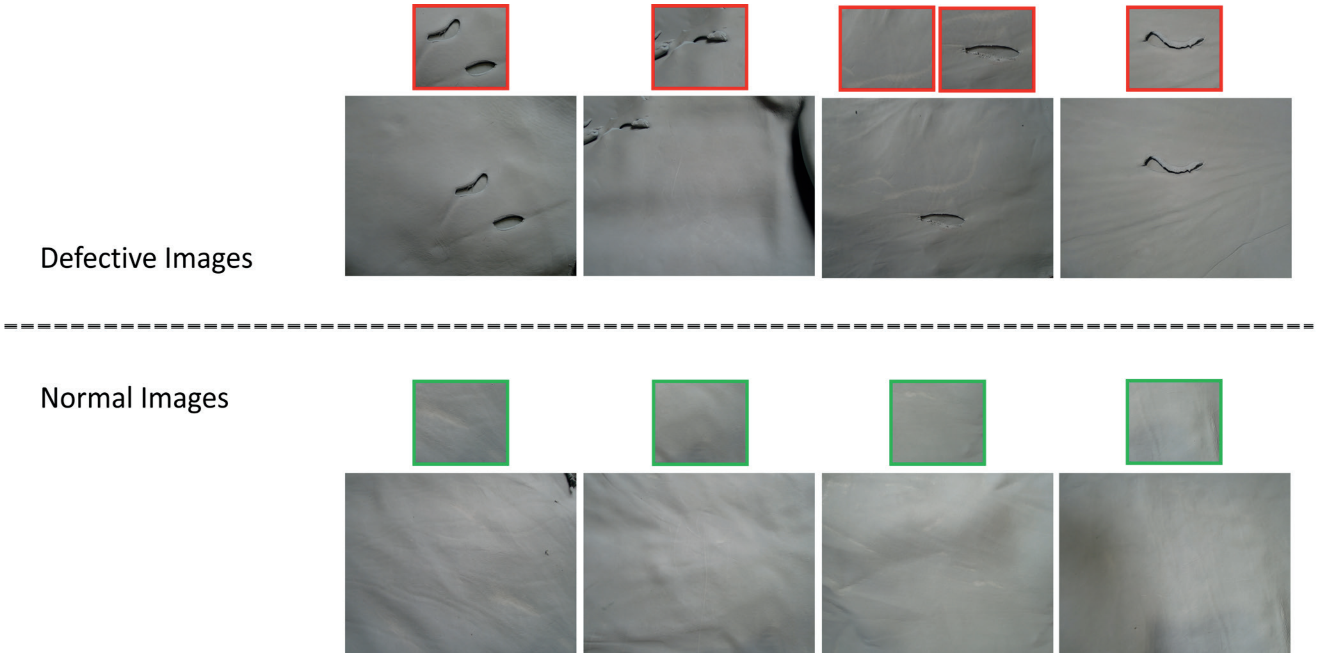

Figure 2. Representative examples of defective and normal images. Image patches on top of the images show abnormal (red boxes) and normal (green boxes) leather surfaces, respectively.

Performance Measures

When performing classification predictions, four types of outcomes could occur:

- **True Positive** (*TP*): When a defected leather sample is predicted as defected by the model,
- **True Negative** (*TN*): When a leather sample without any defects is predicted as non-defected by the model,
- **False Negative** (*FN*): When defects were present in leather but the model predicted it as non-defective; it is also called as a Type 2 error,
- **False Positive** (*FP*): When the leather sample was non-defective but the model predicted it as defective; it is also known as a Type 1 error.

Several different performance measures based on the confusion matrix are employed to assess the classification performance of the methods. These measures include the classification accuracy that measures the percentage of correct predictions

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{8}$$

As per equation (8), it is computed by dividing the number of correct predictions by the number of total predictions. The accuracy measure computed for the training images is termed as the train accuracy. Accuracy for the validation images is the validation accuracy and the accuracy computed on the test image set is known as the test accuracy.

The precision of the classifier quantifies what proportion of positive predictions were deemed correct and is given as

$$\text{Precision} = \frac{TP}{TP+FP} \tag{9}$$

A related measure is recall which measures the proportion of actual positives which were identified correctly

$$\text{Recall} = \frac{TP}{TP+FN} \tag{10}$$

The F1-score is the average of the precision and recall

$$\text{F1} - \text{score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

A Receiver Operating Characteristic (ROC) curve is a fundamental method to evaluate the classifier's test assessment for various classification thresholds. A plot between the sensitivity (true-positive rate) as a function of the specificity (false-positive rate) is shown in this curve for different parameter threshold values. Each point on the graph in the ROC plot is a pair of FPR and TPR values for a particular threshold point. The AUC therefore reflects how a classifier distinguishes between flawed and non-defect leather.

Benchmark Deep Learning Methods

To this end, we employ ResNet, inception-V3 and Inception-ResNet-V2 as benchmark deep learning based methods for comparative evaluation of the proposed method. The VGG16 method is employed as the baseline method for comparison. ResNet has several architectures with different number of layers.[18] For this work, we employed ResNet-50 because for the given amount of data, it gave the best performance. We employed Inception-V3 as it has been used by researchers before as well on related tasks. Inception-ResNet-V2 was implemented as given by C. Szegedy et al.[19] Hyperparameter values used in these algorithms are given in Table III.

State-of-the-art Methods for Comparison

There are only a handful of machine learning approaches reported in literature for wet-blue leather classification. Also, these learning approaches do not publicly share their source code or executable programs that can be used for reproducing their results for comparative evaluation. Therefore, in this work, we compare the performance of the proposed method with contemporary interest point based machine learning techniques and benchmark deep learning methods discussed above. The main motivation for choosing these techniques is their widespread use in literature for similar defect classification problems. We briefly discuss the use of selected techniques in relevant defect detection problems.

Hassanin et al. applied SURF features to classify defects on a printed circuit board (PCB).[20] Zheng D. used Harris corner features for classification of patterns in fabric design.[21] Shang et. al. used inception-v3 with transfer learning to classify and recognize rail surface defects.[22] Wen et. al. used ResNet-50 with transfer learning for fault diagnosis.[23] Shahin et. al. has applied ensemble strategy on skin lesion classification using Inception-v3 and ResNet-50 as an ensemble algorithm.[24]

Therefore, in this work, we compare the classification performance of the proposed CNN based ensemble with the state-of-the-art feature descriptors including SURF by Bay et al.,[20] FAST by Rosten and Drummond[21] and BRISK points proposed by Leutenegger et al.[22] employed in a classification framework. We also employ the well-known Harris corner points algorithm by Harris and Stephens[23] as a baseline descriptor for comparison purpose. These descriptors are employed in a well-known bag-of-keypoints based classification framework[24] (with multiclass SVM as the classifier) for a fair comparison. In addition, we employ ResNet, inception-V3 and Inception-ResNet-V2 as benchmark deep learning based methods and VGG-16 as a baseline deep learning method.

Results

All the algorithms were validated on 328 images during the training process. Before giving the images as an input to the model, all these images were pre-processed. The hyper-parameters used for training are given in Table III. The classification results of the proposed method are presented in Figure 3. It can be observed from the confusion matrices that the proposed Inception-V3+ResNet-50 architecture obtained a high percentage of correct predictions with a high precision and has the ability to robustly identify almost all irregular features in general. This is also confirmed by the ROC curves and the high AUC values as evident in Figure 3. The high AUC values also exhibit the ability of the proposed method to adapt to various applications where different threshold values may be required.

Table III

Hyper-parameter values used in algorithms

Name	Hyper-Parameter Value
Global average pooling	Replaced fully connected layer
Output layer	Activation: Softmax
Number of epochs	20
Batch size	4, 12
Optimization method	SGD & Adam (Learning rate = 0.001)

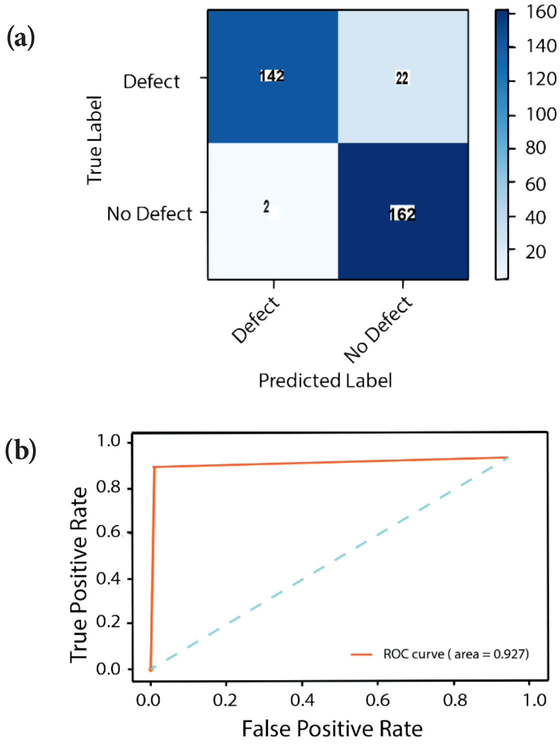

Figure 3. Confusion matrices and ROC curves for the proposed method. **(a)** confusion matrix **(b)** ROC curve

Table IV

Comparison with the state-of-the-art methods in terms of classification accuracy. T Acc stands for Training Accuracy, Val Acc for Validation Accuracy and Test Acc for Test Accuracy.

Serial #	Model	Training Acc (%)	Validation Acc (%)	Test Acc (%)
1	SURF1	94.44	71.67	83.33
2	FAST16	81.67	81.67	73.33
3	BRISK9	94.44	60.00	75.00
4	Harris6	95.00	63.33	76.67
5	InceptionV3+ResNet-50	97.64	96.29	92.68

Comparison with Descriptors Based Machine Learning Methods

In this section, we evaluate the efficacy of the proposed method in comparison to the state-of-the-art feature descriptors based learning methods (discussed in Section 4.4) in terms of classification accuracy, precision, recall, F1-score and AUC. Table IV compares the performance of the proposed methods with the state-of-the-art methods in terms of classification accuracy on the training, validation and test sets. The proposed Inception-V3+ResNet-50 method outperforms all other methods in terms of training accuracy as well as its generalization ability on unseen images. The state-of-the-art methods obtained high training accuracy but in general failed to generalize well on unseen examples in the validation and test sets. The SURF[20] and the FAST[21] methods performed at par on test data with better generalization ability as compared with the BRISK[22] and the Harris[17] descriptors. It is evident from these results that the proposed CNN based method is more suitable as compared with the state-of-the-art machine learning methods due to its robust prediction ability.

It can be observed from Table V that the proposed Inception-V3+ResNet-50 method outperforms all other compared methods in terms of precision, recall, F1-score and AUC. The SURF method

Table V

Comparison of methods in terms of precision, recall, F1-score and AUC.

Serial #	Model	Precision	Recall	F1-score	AUC
1	SURF1	0.86	0.80	0.83	0.90
2	FAST16	0.68	0.87	0.76	0.66
3	BRISK9	0.80	0.67	0.73	0.78
4	Harris6	0.83	0.67	0.74	0.76
5	InceptionV3+ResNet-50	0.93	0.93	0.93	0.927

Table VI

Comparison of models in terms of accuracy, transfer learning, batch normalization, batch size using Global Average Pooling and image size of 500×375.

Serial #	Image Size	Model Name	Train Accuracy (%)	Validation Accuracy (%)	Test Accuracy (%)	Optimizer	GAP	Transfer Learning	Scratch	Batch Normalization	Batch Size	AUC
1	500×375	VGG-16	62.5	0	50	SGD	✓	×	✓	×	4	50
2	500×375	VGG-16	63.08	0	50	Adam	✓	✓	×	×	4	50
3	500×375	VGG-16	67.54	0	50	Adam	✓	×	✓	×	4	50
4	500×375	ResNet-50	99.33	97.31	89.76	Adam	✓	×	✓	✓	4	89.8
5	500×375	ResNet-50	98.08	99.27	89.47	SGD	✓	×	✓	✓	4	89.5
6	500×375	ResNet-50	100	94.65	53.54	Adam	✓	✓	×	×	4	53.5
7	500×375	ResNet-50	100	99.63	90.85	SGD	✓	✓	×	×	4	90.7
8	500×375	Inception-v3	95.48	94.23	65.74	SGD	✓	×	✓	✓	4	65.7
9	500×375	Inception-v3	95.58	98.85	90.15	SGD	✓	✓	×	×	12	90.1
10	500×375	Inception-v3	90.19	40.38	33.85	Adam	✓	×	✓	×	12	33.9
11	500×375	Inception-v3	99.04	97.31	74.4	Adam	✓	✓	×	×	12	74.4
12	500×375	Inception-ResNet-V2	97.88	49.23	24.8	Adam	✓	×	✓	✓	12	24.9
13	500×375	Inception-ResNet-V2	98.56	99.61	74.8	SGD	✓	✓	×	×	12	74.9
14	500×375	Inception-ResNet-V2	99.9	100	54.72	Adam	✓	✓	×	×	4	54.8
15	500×375	Inception-ResNet-V2	97.5	96.15	76.77	Adam	✓	×	✓	×	4	76.8
16	500×375	Inception-V3+Inception-ResNet-V2	99.45	94.84	91.52	Adam	✓	✓	×	×	4	91.5
17	500×375	Inception-V3+ResNet-50	97.64	96.29	92.68	SGD	✓	✓	×	×	4	92.7

generalizes well on the test data as quantified by its AUC which is in agreement with its accuracy. Despite competitive accuracy, the FAST method did not generalize well in terms of AUC. This suggests that although the FAST method can obtain correct predictions, it may not be much more reliable in different applications, where different thresholds are to be set for the classifier. The results of the Harris and the BRISK features are also in line with their accuracy scores.

Comparison with Deep Learning Based Methods

To harness the true performance of the CNN based methods, we experimented with multiple optimization functions and found that the Adam and the Stochastic Gradient Descent (SGD) optimizers are best suited for our problem. We also experimented with transfer learning (with ImageNet[12] weights) in comparison to training from scratch. To aid training from scratch batch normalization was also employed. Some important conclusions drawn from the results in Table V include: 1) The Adam optimizer was found to be the preferred choice for all the networks, 2) Transfer learning from ImageNet weights was found to be more useful in comparison to training from scratch, 3) Batch normalization could not help in improving the results when training the networks from scratch.

Two key observations from our experiments in Table VI are that: 1) knowledge learned from ImageNet transfers better to the task at hand as compared with training from scratch. 2) Inception-V3 and ResNet-50 stand out as compared with other architectures in terms of their performance on leather defect classification. Therefore, when performing our ensembling experiments, we used pre-trained architectures on ImageNet. Also, we select the Inception-V3+ResNet-50 as our proposed ensemble architecture in comparison and compare it with other ensemble approaches. The proposed Inception-V3+ResNet-50 architecture outperformed all other CNN variants in terms of all three accuracy measures when trained using the SGD optimizer in a transfer learning setting. It is also evident that the proposed architecture could be trained with the largest batch size owing to its reduced parameters and compactness. The proposed architecture takes only 1.3 milli seconds to classify an image of resolution 500×375. Surprisingly, VGG-16[25] could not perform well on this task. This result suggests that only sixteen layers of the VGG-16 network are shallow to learn effective representations for the leather classification task. ResNet-50 learned effective representations and exhibited the second best performance in terms of the accuracy measures. Interestingly, it performed slightly better than the Inception-ResNet-V2 architecture in terms of validation and test accuracies.

Class Activations Maps (CAM)

The important region(s) in images utilized by the CNN to predict the class label of an input image can be visualised in several ways, such as gradient descent class activation mappings and global average pooling class activation mappings, etc. In order to interpret the output decision made by any of the CNN architectures investigated in this study, we employ class activation mappings to produce heat maps, which show regions of high importance that influenced the classification output of the method. Figure 4 shows the class activation maps of the proposed method in comparison to other deep learning approaches. In Figure 4, yellow and pink colors are used to represent the regions of high importance according to a particular classifier. Ideally, the regions of high importance should be abnormal surfaces for correct prediction of classes. It is evident from Figure 4 that the proposed method considers the abnormal surfaces (defects) to classify the defective images. Apart from the proposed method only Inception-V3 is able to recognize the abnormal regions to a considerable extent. Otherwise, all other compared methods are not able to focus on important regions potentially leading to misclassifications.

Conclusion

Automated visual inspection of leather in an industrial setting has gained considerable attention recently. Numerous machine learning approaches have been proposed in the past, however, convolutional neural networks based approaches are scarce. In this work, we propose an ensemble convolutional neural network for visual inspection of wet-blue leather. We also present a new dataset of high-resolution leather images. Our proposed technique was able to outperform more than 10 deep learning and machine learning based methods in terms of both test accuracy and AUC score. Our model was able to obtain test accuracy of 92.68% and AUC score of 92.7%. Despite, it's competitive performance, the proposed method would require adaptation for real-time application, which includes fine tuning on video data. In the future, an important direction is to develop a system that can classify leather data in a real-world industrial setting. Another important future direction is to adapt the current system to classify multiple defect types. Finally, the

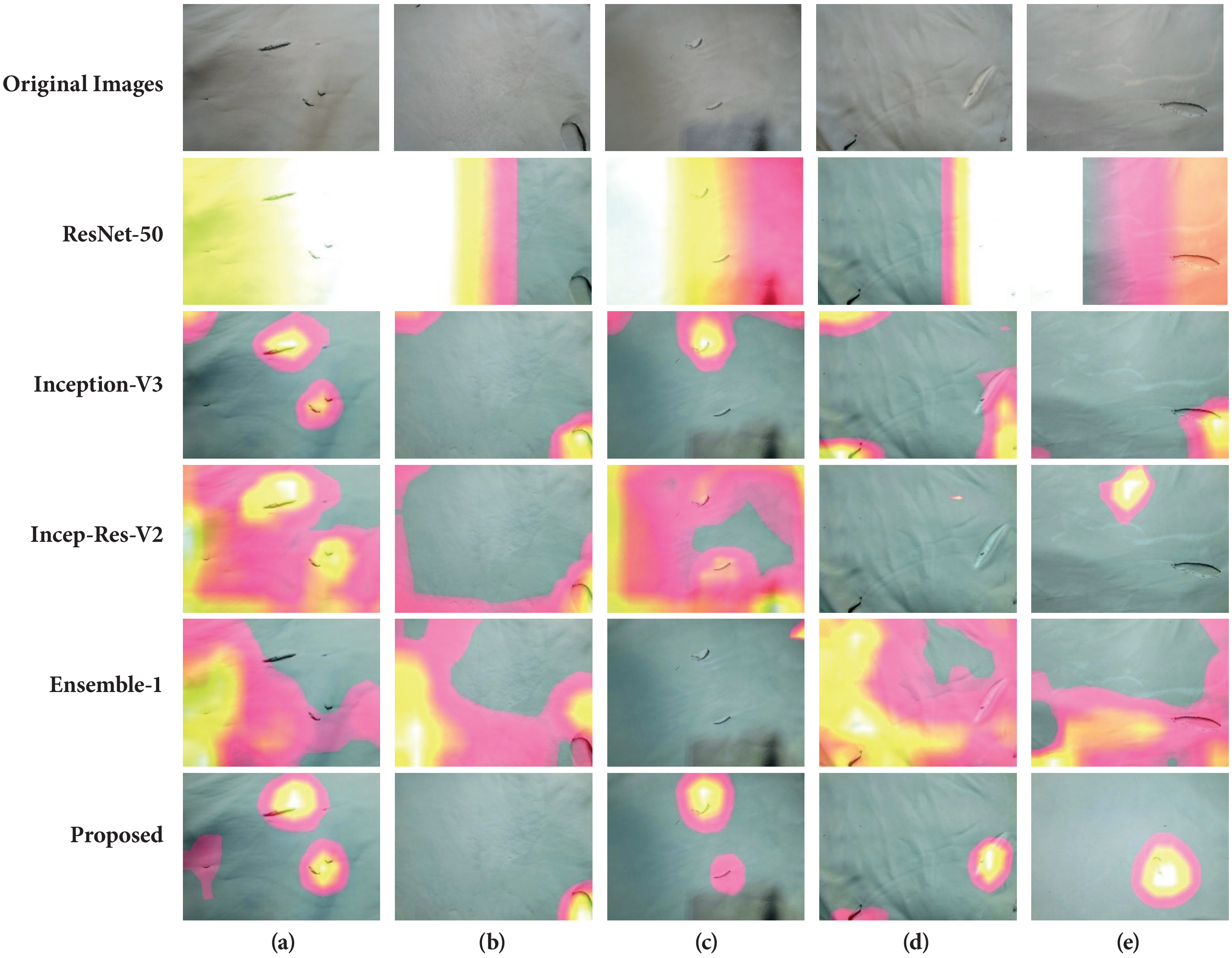

Figure 4. Class activations maps of ResNet-50, Inception-v3, Inception-ResNet-v2 (Incep-Res-V2), Inception-v3 & Inception-ResNet-v4 (Ensemble-1), and Inception-v3 and ResNet-50 (Proposed), respectively.

development of such systems that can characterize various defect types in terms of their properties can potentially lead to artificial intelligence based automated quality grading of leather samples.

Acknowledgment

This work was supported by NZ Leather and Shoe Research Association (LASRA®), Palmerston North, New Zealand through the Ministry of Business, Innovation and Employment (MBIE) grant number LSRX1801.

References

1. C. Koch, K. Georgieva, V. Kasireddy, B. Akinci, and P. Fieguth, "A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure," *Adv. Eng. Informatics*, vol. 29, no. 2, pp. 196–210, 2015, doi: https://doi.org/10.1016/j.aei.2015.01.008.
2. H. Zheng, L. X. Kong, and S. Nahavandi, "Automatic inspection of metallic surface defects using genetic algorithms," *J. Mater. Process. Technol.*, vol. 125–126, pp. 427–433, 2002, doi: https://doi.org/10.1016/S0924-0136(02)00294-7.
3. H. Y. T. Ngan, G. K. H. Pang, and N. H. C. Yung, "Automated fabric defect detection—A review," *Image Vis. Comput.*, vol. 29, no. 7, pp. 442–458, 2011, doi: https://doi.org/10.1016/j.imavis.2011.02.002.
4. S. Vaidya, P. Ambad, and S. Bhosle, "Industry 4.0 – A Glimpse," *Procedia Manuf.*, vol. 20, pp. 233–238, 2018, doi: https://doi.org/10.1016/j.promfg.2018.02.034.
5. M. Egmont-Petersen, D. de Ridder, and H. Handels, "Image processing with neural networks—a review," *Pattern Recognit.*, vol. 35, no. 10, pp. 2279–2301, 2002, doi: https://doi.org/10.1016/S0031-3203(01)00178-9.
6. K. M. C. Mohammed, S. SrinivasKumar, and G. Prasad, "Defective Texture Classification using Optimized Neural Network Structure," *Pattern Recognit. Lett.*, 2020.
7. J. DENG, J. LIU, C. WU, T. A. O. ZHONG, and G. GU, "A Novel Framework for Classifying Leather Surface Defects Based on a Parameter Optimized Residual Network."
8. M. Aslam, T. M. Khan, S. S. Naqvi, G. Holmes, and R. Naffa, "Ensemble Convolutional Neural Networks with Knowledge Transfer for Leather Defect Classification in Industrial Settings," *IEEE Access*, vol. 8, pp. 198600–198614, 2020.
9. H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, pp. 400–407, 1951.
10. D. P. Kingma and J. Ba, "Adam: {A} Method for Stochastic Optimization," 2015, Online. Available: http://arxiv.org/abs/1412.6980.
11. S. Mannor, D. Peleg, and R. Rubinstein, "The cross entropy method for classification," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 561–568.
12. J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
13. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826. https://github.com/keras-team/keras-applications/blob/master/keras_applications/inception_v3.py.
14. M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - Weakly-supervised learning with convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 685–694.
15. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.
16. P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with Convolutional Networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1713–1721.
17. T. Durand, N. Thome, and M. Cord, "WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4743–4752. https://github.com/keras-team/keras-applications/blob/master/keras_applications/vgg16.py.
18. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. https://github.com/keras-team/keras-applications/blob/master/keras_applications/resnet50.py.
19. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the Impact of Residual Connections on Learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 4278–4284. https://github.com/keras-team/keras-applications/blob/master/keras_applications/inception_resnet_v2.py.
20. H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Comput. Vis. Image Underst.*, vol. 110, no. 3, pp. 346–359, 2008.
21. E. Rosten and T. Drummond, "Fusing points and lines for high performance tracking," in *Tenth IEEE International Conference on Computer Vision (ICCV'05)*, 2005, vol. 2, pp. 1508–1515.
22. S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary Robust invariant scalable keypoints," in *2011 International Conference on Computer Vision*, 2011, pp. 2548–2555, doi: 10.1109/ICCV.2011.6126542.
23. C. Harris and M. Stephens, "A combined corner and edge detector," in *In Proc. of Fourth Alvey Vision Conference*, 1988, pp. 147–151.
24. G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.
25. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2015.

Lifelines

Qian Zhang received her Master's degree in 2020 from School of Materials Science and Engineering at Zhengzhou University, Zhengzhou, China, majoring in leather chemistry and engineering. Her research focuses on the clean production in leather making.

Jie Liu received his PhD degree in 2007 from Institute of Mechanics, Chinese Academy of Sciences, Beijing, China. He currently is an associate professor at School of Materials Science and Engineering at Zhengzhou University, Zhengzhou, China. From 2016 to 2017, he worked as a visiting scientist at ERRC, USDA in Cheng-Kung Liu's group. His current research interests focus on green composite materials based on natural polymers and their applications in packaging, biomedical and environmental fields.

Xiumin Li received his PhD degree in 2017 from Hirosaki University, Aomori, Japan. He is working for Zhengzhou University as an associate professor at School of Materials Science and Engineering, Zhengzhou, China. His research interests include the electro-catalyst materials and nanostructure engineering. He has published more than 50 papers and co-authored 1 book in the field of nanostructured material.

Hui Liu received his PhD degree in 2020 from Zhengzhou University, Zhengzhou, China. His research focuses on the mechanism and application of simplified clean production of ecological leather with biological enzymes.

Yadi Hu is a PhD candidate of School of Materials Science and Engineering at Zhengzhou University, majoring in the materials physics and chemistry. Her research focuses on the degradation mechanism of modern and historical leathers under the effect of storage environmental conditions.

Keyong Tang received his PhD degree in 1998 from Sichuan University, Chengdu, China. He is working for Zhengzhou University as a professor at School of Materials Science and Engineering, Zhengzhou, China. His research interests include the leather structure and properties. He has published more than 100 papers, co-authored 4 books and edited 1 book in the field of leather chemistry and engineering.

Xueru Guo received her B.S. degree in Environmental Engineering in Shaanxi University of Science & Technology in 2017. Now she is undertaking her M.D. degree in Biomass Chemistry and Engineering at Sichuan University. Her research focuses on chrome-free tanning technology.

Yue Yu, see *JALCA* 115, 190, 2020.

Ya-nan Wang, see *JALCA* 112, 258, 2017.

Bi Shi, see *JALCA* 99, 220, 2004.

Qiao Xia received his Bachelor's degree from Chengdu University of TCM 2019. He is currently studying for a master's degree with Prof. Zongcai Zhang at Sichuan University. His research interests are in leather chemistry and engineering.

Zongcai Zhang received his Doctor's degree in leather chemistry and engineering from Sichuan University in 2005. His research mainly focuses on the cleaner technology and development of fine chemicals for leather and fur.

Meina Zhang received her Master's degree from Sichuan University 2020.He is currently studying for a Doctor's degree at University of Liverpool.Her research concentration is in chemistry.

Yingxuan Wang received her Bachelor's degree from Xi'an University of Architecture and Technology 2019. She is currently studying for a Master's degree with Prof. Zongcai Zhang at Sichuan University.

Hong Dai received her doctorate degree in leather chemistry and engineering from Sichuan University in 2006. Her research mainly focuses on analysis and testing of leather and fur.

Masood Aslam received the B.S. degree in electrical engineering from The University of Faisalabad, Faisalabad, Pakistan, in 2013, and the M.S. degree in electrical engineering from FASTNUCES, Islamabad, Pakistan, in 2018. He is currently working as Research Associate with the Visual Computing Technology (VC-Tech) Lab, Islamabad. Before joining VC-Tech, he worked as a Research Assistant with FAST-NUCES. His research interests include computer vision and image processing.

Tariq M. Khan (Member, IEEE) received the Ph.D. degree from Macquarie University, Sydney, Australia. He is currently working as a Research Fellow at Deakin University. At CUI, he also serving as a Head of the Image and Video Processing Research Group (IVPRG). He has over ten years of research, and teaching experience in universities including CUI, Macquarie University, and UET Taxila, Pakistan. He was a recipient of several awards including iMQRES scholarship, Macquarie University Postgraduate Research Student Support (PRSS) travel grants and SRGP. He is interested in both digital image processing (with an emphasis on medical imaging) and machine learning. His research interests include most aspects of image enhancement, pattern recognition, and image analysis.

He has published 40 journals and 22 conference papers in well-reputed journals and conferences e.g. IEEE Transactions on Image Processing, Pattern Recognition, IEEE Access, Expert Systems with Applications, JRTIP, ICIAR, VCIP, and DICTA. His research has received research funding from Effat University, HEC Pakistan, MBIE New Zealand, and KSA International Collaboration Grant over $1M in past years.

Syed Saud Naqvi received the B.Sc. degree in computer engineering from the COMSATS Institute of Information Technology, Islamabad, Pakistan, and the M.Sc. degree in electronic engineering from the University of Sheffield, U.K., in 2005 and 2007, respectively, and the Ph.D. degree from the School of Engineering and Computer Science, Victoria University of Wellington, New Zealand, in 2016. He is currently working as an Assistant Professor with the COMSATS Institute of Information Technology Islamabad, Pakistan. His research interests include saliency modeling, medical image analysis, scene understanding, and deep learning methods for image analysis.

Geoff Holmes received the degree in applied science from Kingston University, in 1986, and the H.N.D. degree in leather technology from the University of Northampton's Institute for Creative Leather Technologies, in 1989. He is a Leather Technologist. He has over 30 years' experience in the leather industry in various production and technical roles, and for the last 14 years has been involved in research at the New Zealand Leather and Shoe Research Association (LASRA). He is the current LASRA Director.

Rafea Naffa received the Ph.D. degree in biochemistry from the School of Fundamental Sciences, Massey University, in 2017. He is currently a Scientist with the Leather and Shoe Research Association, Palmerston North, New Zealand. He has published more than 18 peer-reviewed articles since 2017, including six method articles. He is also studying the extraction, purification, and characterization of collagen from different animal species for several applications. His research focuses on the analysis of collagen in skins and hides where he developed several analytical methods for the separation and quantitation of collagen crosslinks and advanced glycation end products using LC-MS.