IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 3, 2019, accepted November 26, 2019, date of publication December 3, 2019, date of current version December 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957427

On the Application of Automated Machine Vision for Leather Defect Inspection and Grading: A Survey

MASOOD ASLAM[1], TARIQ M. KHAN[1], (Member, IEEE), SYED SAUD NAQVI[1], GEOFF HOLMES[2], AND RAFEA NAFFA[2]

[1]Department of Electrical and Computer Engineering, COMSATS University Islamabad, Islamabad Campus, Islamabad 45550, Pakistan
[2]NZ Leather and Shoe Research Association (LASRA), Palmerston North 4414, New Zealand

Corresponding author: Masood Aslam (masood.aslam2011@gmail.com)

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

ABSTRACT Reliably and effectively detecting and classifying leather surface defects is of great importance to tanneries and industries that use leather as a major raw material such as leather footwear and handbag manufacturers. This paper presents a detailed and methodical review of the leather surface defects, their effects on leather quality grading and automated visual inspection methods for leather defect inspection. A detailed review of inspection methods based on leather defect detection using image analysis methods is presented, which are usually classified as heuristic or basic machine learning based methods. Due to the recent success of deep learning methods in various related fields, various architectures of deep learning are discussed that are tailored to image classification, detection, and segmentation. In general, visual inspection applications, where recent CNN architectures are classified, compared, and a detailed review is subsequently presented on the role of deep learning methods in leather defect detection. Finally, research guidelines are presented to fellow researchers regarding data augmentation, leather quality quantification, and simultaneous defect inspection methods, which need to be investigated in the future to make progress in this crucial area of research.

INDEX TERMS Leather defects, segmentation, classification, machine learning, computer vision.

I. INTRODUCTION

Millions of tons of hides and skins are generated as co-product from the slaughtering of animals for their meat each year. They are mostly converted into leather, the most important economic by-product of the meat industry. In 2003, the global leather industry produced approximately 18 billion ft^2, with an estimated value of US$ 40 billion [1]. Developing countries now produce more than 60% of the leather requirements world-wide. New Zealand hides and skins, especially herd skins, make a major contribution to leather worldwide by providing raw skins for the tanning industry [2]. Skins are mostly sourced from sheep, cow, deer, and goat in New Zealand. In 2011, 75% of sheep and lambskin were exported, mainly to the garment industry [2].

The associate editor coordinating the review of this manuscript and approving it for publication was An-An Liu.

Before leather can be exported as a manufactured product, it must undergo numerous processing steps. The general leather processing steps include preparation for tanning, tanning, and finishing. These processing steps have multiple stages that depend upon the type of material used and the kind of leather required as a product. The processed leather is then subjected to leather quality grading, which is the process of categorizing leather based on the surface defects found during the inspection. The high demands for quality assurance are driven by global customer requirements, and increasingly rejection costs. Accordingly, the inspection of skin surface defects is essential for objectivity and reliability. Currently, the process of surface defect inspection and grading is carried out by human inspectors. The large scale of leather production makes defect inspection a labor-intensive and time-consuming process, which can be a potential bottleneck in the production pipeline,

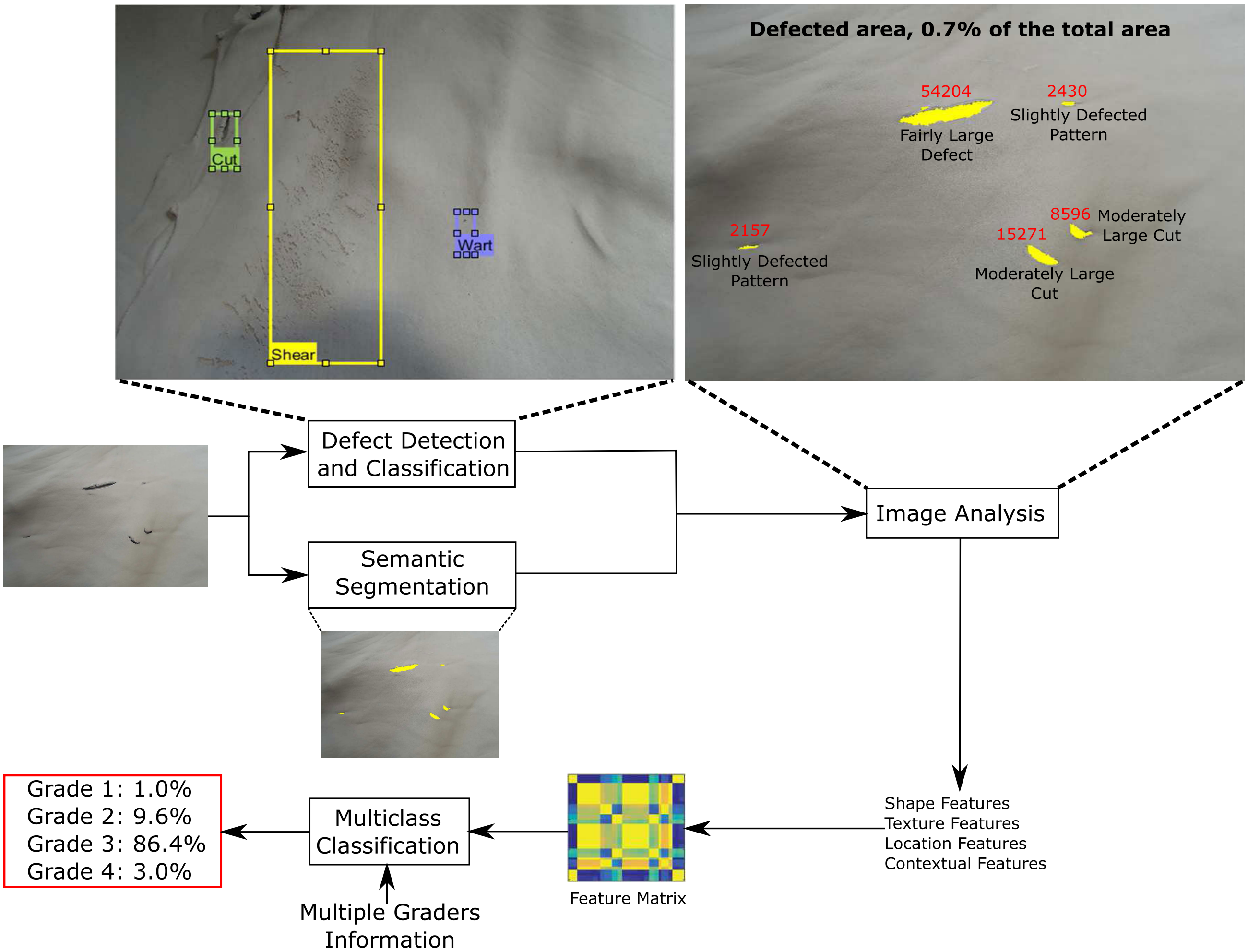

FIGURE 1. Overall pipeline for leather visual defect inspection - a guideline for machine vision systems.

thus motivating the design and development of automated visual inspection systems for surface defect detection and grading.

This decade has seen astounding progress in the application of intelligent systems to real-world problems in areas including but not limited to medicine, telecommunications, finance, medical diagnosis, transportation, information retrieval, energy and many more. The urge for automation has revolutionized the industry sector with expert and intelligent systems finding applications in almost all kinds of industrial processing, ranging from resource optimization to industrial inspection. Intelligent machine vision systems have been at the heart of industrial inspection and surveillance for the past two decades. Image analysis based methods proposed for industrial inspection include both heuristic and machine learning methods. Despite being an important subject in industrial inspection, leather defect inspection has not received much attention yet. The majority of methods that exist for visual defect inspection of leather are heuristic with only limited studies that explore machine learning options for robust performance. Leather quality grading based on defect inspection is an important area of research and rapid advancements in intelligent systems for automatic leather grading are expected in the near future.

Advancements in expert deep learning based systems and their ability to surpass human performance has increased their application in numerous fields including healthcare, automotive industry, telecommunications, industrial visual inspection and many more during recent years. Most convolutional neural networks (CNN) based methods proposed for various computer vision applications can be categorized by one of the following tasks: image classification, detection, semantic segmentation and multiclass classification. Owing to their success in the aforementioned tasks and their ability to match or surpass human level performance, we present a complete deep networks based machine vision pipeline for visual defect grading, which can be utilized by future researchers as a guideline. Figure 1 shows the recommended pipeline for leather defect inspection.

In the first stage defect detection and classification is performed in parallel with defect segmentation. Based on the information from the detection, classification and segmentation; important features for grading including shape, texture, location and context are computed using an image analysis module. This in-depth analysis of the leather defect's characteristics is then fed to a multiclass classifier as a feature matrix to obtain the final leather quality grade. We expand on this pipeline in Section V, discuss and recommend CNN architectures that might be worth investigating for leather defect detection, classification, segmentation and multiclass classification (grading). These guidelines would help researchers in this field to investigate machine vision systems for leather defect detection.

This paper aims at reviewing the various types of leather defects, their effects on leather quality, and progress/developments in the field of visual inspection based quality assessment of leather. We intend to review the recent image analysis based methods for autonomous leather defect identification and discuss their pros and cons. Moreover, motivated by the success of the recent deep learning based methods for autonomous visual inspection in related applications, this paper provides detailed recommendations/guidelines for the design and development of a machine vision based leather defect detection system. We discuss the challenges in the design and development of deep learning methods for leather defect inspection and provide future research directions. This paper also opens avenues for further research in this domain and pose open research questions for researchers pursuing this area. The following are the contributions of this work:

- Develop and introduce a comprehensive nomenclature of leather defects, their characterization and leather quality grading.
- Provide a comprehensive review of image analysis based methods for autonomous leather defect inspection.
- Propose a complete leather defect inspection pipeline using CNNs for defect detection, classification, segmentation and grading.
- Discuss challenges and opportunities in the design and development of deep learning based methods for automatic leather inspection.
- Highlight open problems and future research directions for research and development of machine vision systems for autonomous leather quality grading.

The rest of the paper is organized as follows. Section II provides background knowledge on leather processing and leather quality grading followed by a comprehensive review of leather defects. Section III reviews image analysis based methods for leather defect inspection and their comparison is presented in Section IV. Section V, discusses the application of expert deep learning-based methods for various applications including image classification and segmentation, deep learning-based methods for visual defect inspection. Section VI and VII present information regarding the datasets for leather images and the evaluation measures employed for evaluation of automated inspection systems. Finally, Section VIII discusses the challenges, opportunities and future research directions followed by the conclusions.

II. BACKGROUND KNOWLEDGE

A. LEATHER QUALITY

Leather & leather products have been used by people of all kinds from ancient times. Due to extensive utilization and industrialization, developing countries today account for 60% of leather production in the world, and this share is growing [3]. The leather industry utilizes a meat industry's by-product. Animal skins and hides are processed to produce leather of fine quality. Due to the intrinsic variation of the natural raw material, the skin quality varies significantly. The evaluation of an effective cut-off value that can be obtained from the skin when used in the manufacturing of a product is made based on defects on the skin substrate.

In a changing global scenario, the entire manufacturing industry constantly demands higher product quality and improved productivity to meet both customer requirements and to reduce rejection costs. The increased demands for objectivity, reliability, and efficiency have required the addition of automatic inspection systems in the traditional leather industry. Surface defects reduce the quality and value of the skins. In general, defects of hides and skins are widely classified as antemortem (before animal death), post-mortem (fault after animal death) and as processing defects. Possible defects that occur during the lifetime of the animals are brand marks, tick marks, pox marks, insect bites, wounds, scratches, growth marks, etc.

For the manufacturing of leather products, that mostly require high quality automatic visual inspection of the skin surface, defects are important. Currently, these operations are carried out by human inspectors who tend to miss a significant number of defects because people are mostly unsuited to such a simple and repetitive task. In addition, manual inspection is a labor-intensive task and is very slow; mostly the human inspectors become a critical bottleneck throughout the production process. Hence, it is vital to develop an automated visual inspection system to harmonize the quality assessment process.

Although automated control systems have successfully replaced manual systems in many areas, improving process accuracy is still necessary to reduce false positives (products classified as good if they are defective), false negatives (products classified as defective when they are good) and processing time.

The quality of raw skins and hides depends largely on the amount of antemortem and post-mortem damage. Ante-mortem or post-mortem defects affect the quality of the raw material. Although post mortem defects can be controlled to a certain amount, antemortem damage poses major challenges for the tanner. The quality of raw skin and hide is crucial to the quality of the leather formed and accounts for about 50% to 70% of the production cost, the raw material being the most valuable and most important element in production.

B. LEATHER DEFECTS

Leather defects can be generally classified into natural, mechanical, flaying, curing and tanning defects. The specific

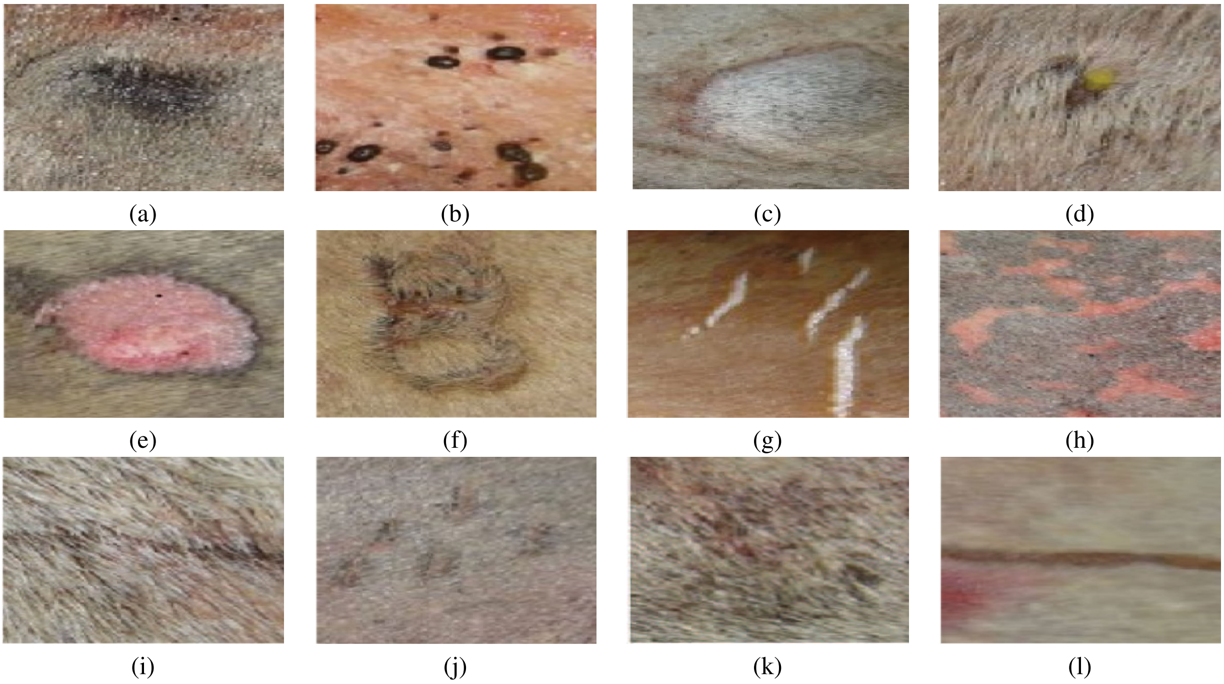

FIGURE 2. Examples of defects in raw hides: (a) bot fly closed wounds, (b) ticks marks, (c) vaccine abscess, (d) bot fly open wounds, (e) flay mark, (f) brand marks made from hot iron, (g) wrinkles, (h) photosensibility, (i) closed cuts, (j) scabies, (k) horn fly wounds and (l) open cuts.

TABLE 1. Type of defects and their causes.

Defects	Causes
Natural Defects	diseases and parasites on the living animal i.e. anthrax, scars, ringworm, tumours, or lice, warble and ticks etc
Mechanical Defects	living animal such as brands, bruises, scratches, wounds, wire damage etc.
Flaying Defects	scores, cuts, holes, corduroy, pulling machine damage, grain cracks etc.
Curing and storing defects	red heat, salt stains, putrefaction etc.
Tanning Defects	(1) mechanical defects fleshing cuts, uneven thickness due to splitting or having, sammying folds (2) process and chemical defects chrome or salt stains, incorrect moisture content, incorrect physical or chemical properties as specified by ISO norms for wet-blue (or other relevant and agreed upon national or international technical specifications).

causes and characteristics of these defects are listed in Table 1. The tanning defects can further be categorized into mechanical or process and chemical defects. Figures 2 and 3 show representative examples of the aforementioned defect types. These defects are characterized by variety of forms, colors, and textures that emerge from several significant defects in raw hides and leather, which reduce their market value. Images of raw hides with defects taken after skinning and before tanning are shown in Figure 2. Figure 3 shows defected leather samples after the first phase of tanning known as wet blue leather.

A standard system must be in place in the leather industry to determine the quality of the hide or skin. This refers directly to the characteristics of leather in terms of its yield and quality. This standard system is crucial for both the vendor and the buyer. This system is based on different grades of quality and weight range. In some nations, first to third grades are only available, but in some other nations, fourth grade is also available. In general, the grading principle is the same for all countries, given that there will be a slight variation from country to country. The sub-divisions of a hide are given in Figure 4. For this standard, faults and defects are distributed in five groups given in Table 1.

Standard grading norms which are widely accepted by the leather industry for determining leather quality are described below[1]:

First Grade: First grade defects are given in Figure 5, The first grade shall be done according to the following requirements.

[1]The standards for these gradings are taken from https://www.allpi.int/courses-and-publications/reports/manuals/hides-and-skins-improvement published by Africa Leather and Leather Products Institute on 10 October 2016.

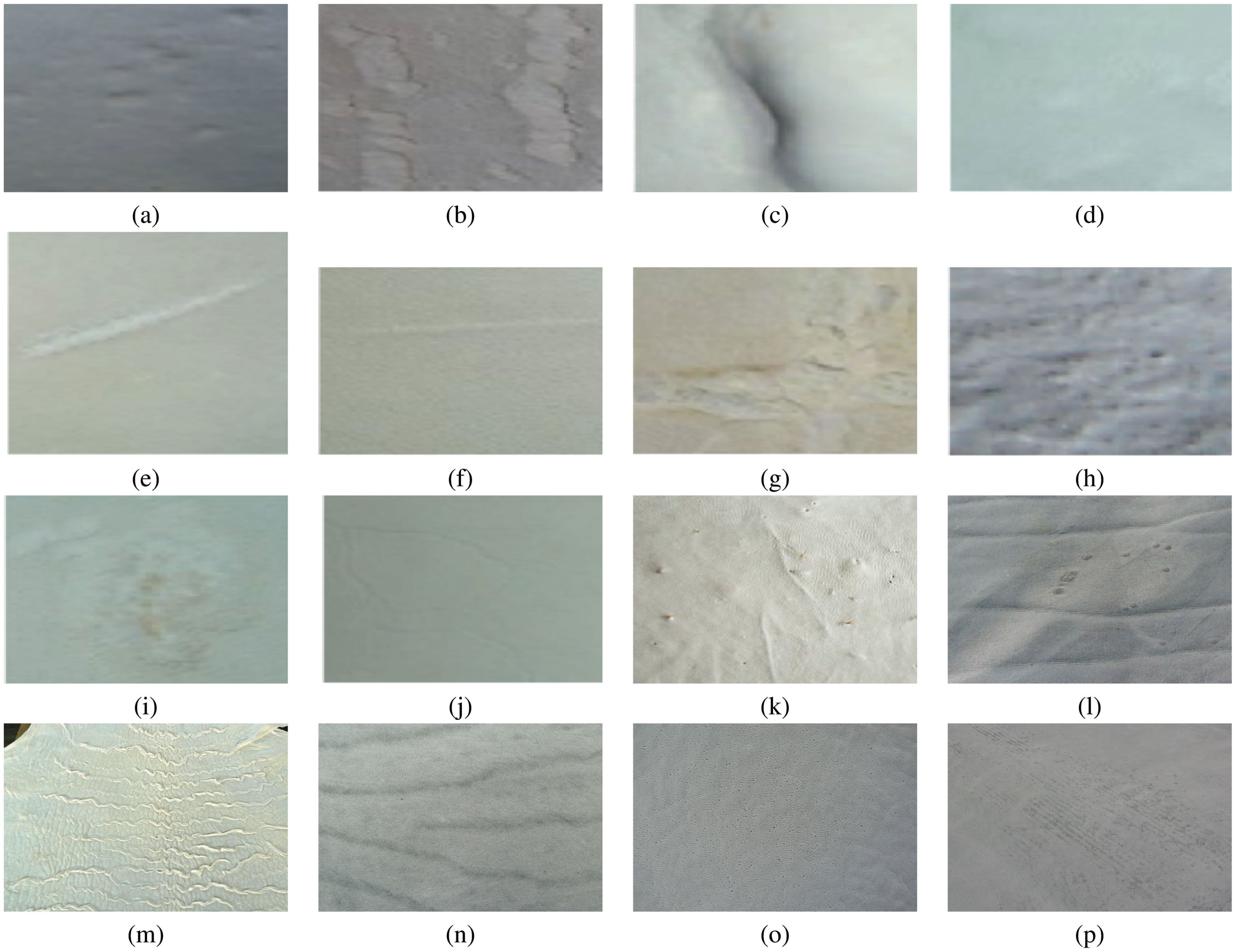

FIGURE 3. Examples of defects in wet blue leather: (a) ticks marks, (b) wrinkles, (c) vaccine abscess, (d) bot fly wounds, (e) open cuts, (f) closed cuts, (g) brand marks made from hot iron, (h) Haematobia irritans wounds, (i) scabies, (j) veining, (k) seed on pp, (l) warts on neck, (m) lap rib on pickled pelt, (n) mild raised rib, (o) pinholes and (p) whitespot shear pattern.

- Hide of the fine pattern, tidy and well-cured, no sign of putrefaction
- Free of defects in the butt and neck region, except for a maximum of 5 blind warbles,
- With just a few score marks or a hole in the bellies,
- Without brand marks.

Second Grade: Second grade defects are given in Figure 6. The second grade shall be done according to the following requirements.

- Hide of the fine pattern, tidy and well-cured, no sign of putrefaction,
- With a few tiny holes or cuts or other groups one and two defects in the butt,
- With a moderate amount of defects from group one, and group two - except for brands and four in the belly and neck,
- With a brand mark that is completely within 18 cm of the perimeter of the hide,
- With a maximum of 10 open warbles or 20 blind warbles,
- With dung and urine stain not more than an area of 30 × 30 cm on each of the hind shanks.

Third Grade: Third grade defects are given in Figure 7. The third grade shall be done according to the following requirements.

- Hide of poor pattern or spoiled,
- some putrefaction defects,
- With group one, two faults - except for brands and four to 30% of the hide region,
- With a brand that has a part of more than 18 cm from the perimeter of the hide,
- With more than 10 open warbles and 20 blind warbles,
- With more stain during and urine than acceptable for grade 2.

Fourth Grade: Fourth grade defects are given in Figure 8. The Fourth grade shall be made according to the following requirements.

- Very bad pattern, very spoiled hide.
- With any kind of defect covering up to 50% of the hide region.

TABLE 2. Summary of image analysis based methods. These methods belong to the general categories of texture-based methods, wavelet-based methods, classifier based techniques, and miscellaneous methods. The techniques are generally classified as belonging to image processing, machine learning or neural network domain. A summary of achieved outcomes is also reported.

Sr. #	Author Name	Description	Technique	Outcomes
1	A.Learch et al. [19]	Processing and segmentation of images of defects and areas of different quality, usually manually marked	Image Processing	line correction for automatic cutting of hides
2	R. Rodrigues et al. [13]	Support Vector Machine for leather defect classification. Compared LIBSVM and Weka SMO techniques.	Machine Learning	accuracy of 99.59% obtained using SVM classifier
3	C. Yeh et al. [14]	Used an algorithm which gave deductible area, unusable area, deviation rate and error rate of each transaction.	Image Processing	obtained an error rate of 1.16%
4	H. Pistori et al. [18]	Used Attribute reduction on 4 different classifiers. Compared these classifiers kNN, C4.5, Naïve Bayes and SVM	Machine Learning	KNN obtained accuracy of 92.3%
5	Fuqiang He et al. [9]	Wavelet transform for Automatic Band selection scheme	Image Processing	normalized energy of 96.2% was achieved
6	V.K. Sahu et al. [25]	Auto adaptive edge detection algorithm. Compared their technique with Canny, Sobel, Prewitts and Roberts	Image Processing	Gives clear and continuous edges as compared to other algorithms
7	K. Vani et al. [10]	Wavelet feature Extraction method. WCF and WSF were used as feature extractor with SVM classifier	Machine Learning	SVM performed better when both WCF and WSF were used
8	H.Bin et al. [15]	Used Feed Forward Neural Network technique. Applied Decision Tree for classification	Neural Network	neural network obtained an accuracy of 90%
9	P. Villar et al. [16]	Used multi-layer perception to train the neural network	Neural Network	MLP achieved an accuracy of 95%
10	C. Ho et al. [26]	Multi-camera fusion based algorithm for marking of leather defects.	Image Processing	percentage error improvement from 1.1% to 0.1%.
11	R.F. Pereira et al. [17]	Used classifiers KNN, MLP and SVM to detect 10 kind of defects.	Machine Learning	SVM with RBF kernel was found to outperform other methods with 86% accuracy
12	Pistori, H et. al [12]	The performance comparison for several classifiers was applied to both raw hide and wet blue leather samples.	Machine Learning	accuracy = 100% when image size was 40x40
13	Limas-Serafim [31]	used multi-resolution pyramids for segmenting the scar and small vein defects in calf leather.	Image Processing	only a few representative visual results on calf leather were presented
14	Lovergine et al [5]	classify the input leather surface types based on gradient orientation and local coherence	Image Processing	only few visual results presented
15	J.L. Sobral et. al [8]	wavelets technique, which uses a bank of optimized filters, where each filter is adjusted to a certain defect.	Image Processing	real time performance
16	Yeh and Perng [22]	semi-automatic methods for wet blue leather defects extraction and classification.	Image Processing	fully quantified grading system
17	Kwak, C. et. al [23]	use of geometric and statistic descriptors, in addition to the use of decision trees for the classification of the leather surface.	Neural Network	accuracy of 91.25% on 140 defect samples

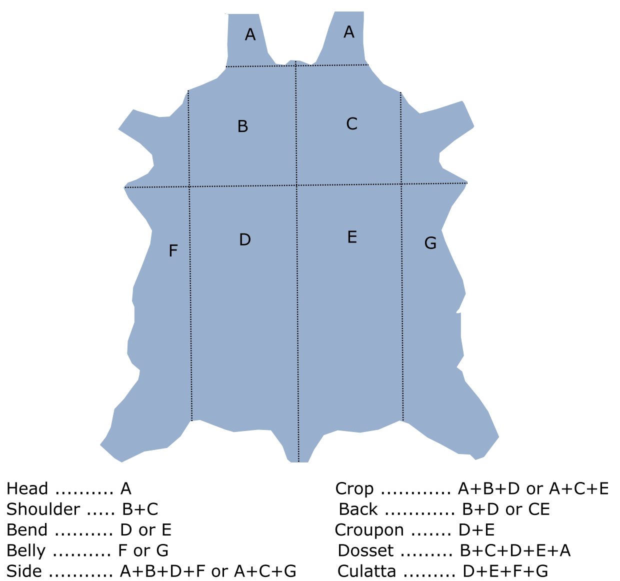

FIGURE 4. Subdivisions of a hide.

III. IMAGE ANALYSIS BASED METHODS FOR LEATHER DEFECT INSPECTION

In past years, a wide variety of image analysis based methods for automated leather defect inspection have been reported in literature. These methods can be broadly classified into texture based methods, Wavelet based methods and classifier based methods. Other prominent methods investigated for leather defect detection include methods based on geometric and statistical descriptors, histogram based methods, edge detection based methods, multi-camera fusion based techniques and semi-interactive methods. All the aforementioned methods are described here in this section and a quantitative comparison of these techniques is presented in the following section. A summary of these techniques is presented in Table 2.

A. TEXTURE BASED METHODS

Limas-Serafim used multi-resolution pyramids to create a leather surface texture model [4]. This method was implemented and applied in calf leather to segment the scar and small vein defects. Lovergine *et al.* used a leather patch as a unit for the classification and input of leather surface types based on local consistency and gradient orientation [5]. This also showed that leather defects such as folds and scars could be detected by segmenting the leather surface's oriented texture map. Branca *et al.* identified leather surface defects according to oriented structures using human vision [6]. The neural network approach was then used to evaluate the texture-oriented approach. The resulting system is flexible and does not rely on the composition, color or size of the defects.

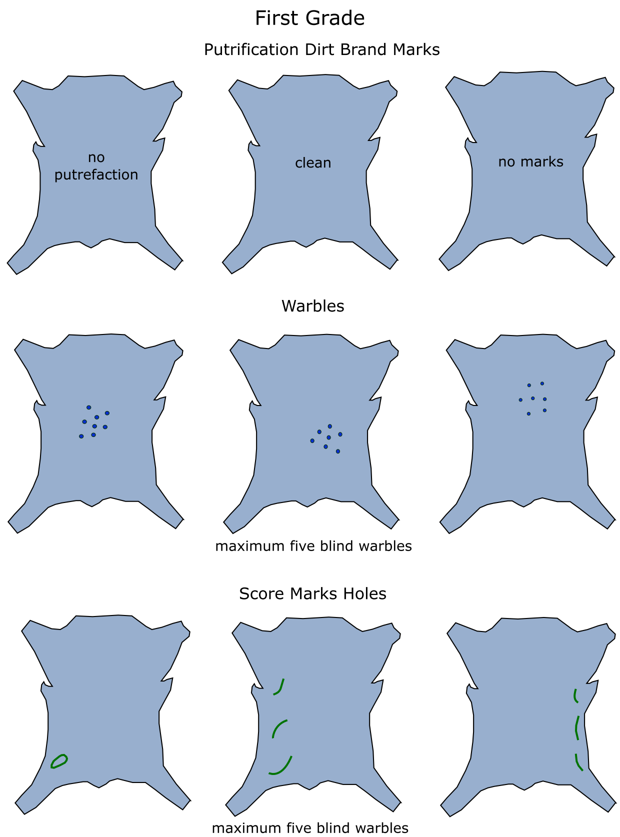

FIGURE 5. **First grade defects on wet blue leather.**

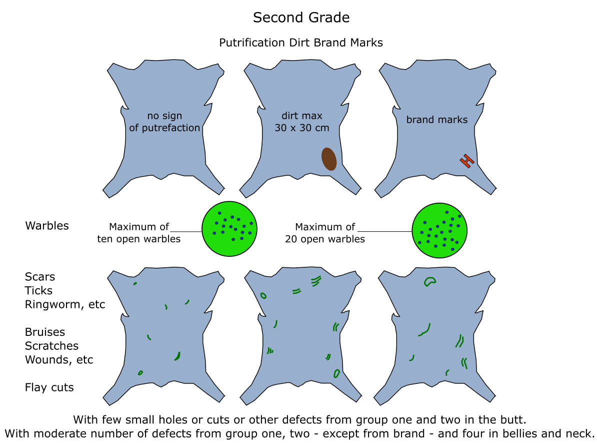

FIGURE 6. **Second grade defects on wet blue leather.**

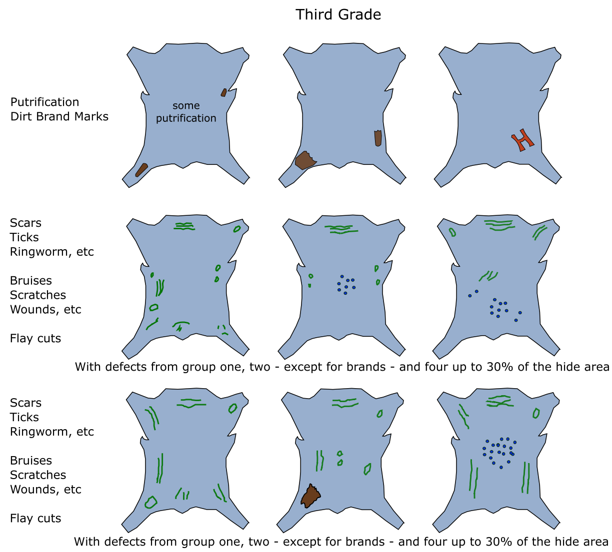

FIGURE 7. **Third grade defects on wet blue leather.**

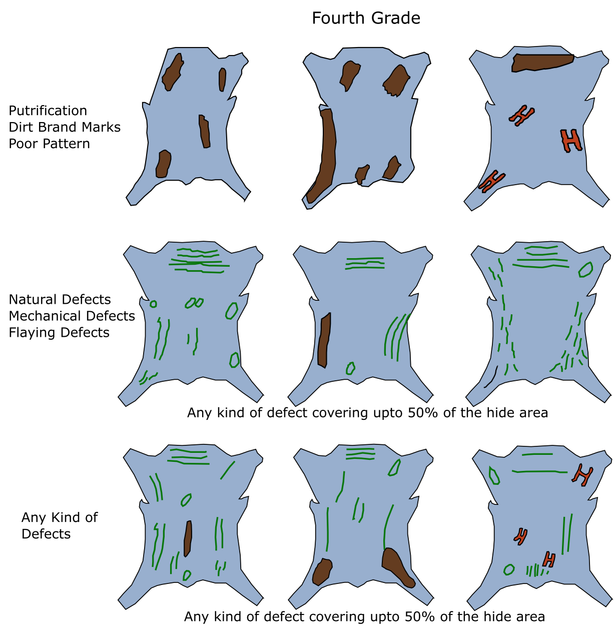

FIGURE 8. **Fourth grade defects on wet blue leather.**

The work of Branca *et al.* relies on orientation flow field modeling for texture analysis [5]. The oriented texture field as a vector can be defined by its components, i.e. the dominant gradient direction and the coherence. The gradient direction can be generally defined as:

$$\theta = \frac{1}{2}\arctan\frac{\sum_{m,n} r_{m,n}^2 \sin 2\theta_{m,n}}{\sum_{m,n} r_{m,n}^2 \cos 2\theta_{m,n}}, \tag{1}$$

where $\theta_{m,n} \in \left[-\frac{\pi}{2}, \frac{pi}{2}\right]$ and the representation in 1 is in polar form. The coherence measure ϵ determines the accord of local gradient orientation in a neighborhood (given by x and y) and is defined as:

$$\epsilon_{m,n} = \nabla_{m,n}\frac{\sum_{x,y} \nabla_{x,y} \cos(\theta_{x,y} - \theta_{m,n})}{\sum_{x,y} \nabla_{x,y}}, \tag{2}$$

where $\nabla_{m,n} = (\nabla_m, \nabla_n)^T = r_{m,n}\exp^{i\theta_{m,n}}$ and as evident from equation 2, ϵ is directly proportional to the closeness of orientations in the neighborhood, given by x and y.

In other methods Kumar and Pang rely on Gabor filters to mimic the early human visual system [7]. The idea is to characterize the textures of regions using the various orientations detected by the Gabor filters. The general 2D Gabor function is expressed as:

$$\psi(x,y) = \frac{1}{2\pi\gamma\eta} \exp\left[-\frac{1}{2}\left(\frac{x^2}{\gamma^2}+\frac{y^2}{\eta^2}\right)\right] \exp(2\pi jfx), \quad (3)$$

where f is the frequency of the filter, γ and η represent the Gaussian envelope along the major and minor axis.

B. WAVELET BASED METHODS

Sobral *et al.* introduced a new wavelet-based method using optimized filter banks where each filter is adjusted to a particular defect [8]. These filters and the wavelet ranges are selected based on the maximization of attributes obtained from leather flaws. Such a method can detect faults even if there is a small change in attributes and it is based on Gabor filters as described by [7]. Furthermore, it has been shown that this technique is fast enough to detect in real-time. He *et al.* automatic band selection scheme was developed based on wavelet transform and wavelet energy coefficients distributed in different frequency channels, to determine decomposed sub-images and the number of multi-resolution levels [9]. They achieved normalized energy of 96.2% and claimed real-time performance. Jawahar *et al.* gave a feature extraction method for leather defects identification using wavelet feature extraction [10]. Wavelet Co-occurrence Feature (WCF) and Wavelet Statistical Feature (WSF) were used as feature extractors. They used SVM as the classifier. They claimed to get good results when both WCF and WSF were combined. Their system was able to discriminate between defective and non-defective leather.

C. CLASSIFIER BASED SYSTEMS

Many classifier-based systems have been proposed for leather defect detection in literature. Most of them employ a definitive set of features that can be generally classified as first-order statistical measures, second-order statistical measures, spectral measures or image-level descriptors (local binary patterns and Gabor features). A set of measures based on the co-occurrence matrix [11] that are found common among several implementations are energy E, entropy S, inertia I and homogeneity H, and are defined as:

$$E = \sum_i \sum_j C(i,j)^2, \quad (4)$$

$$S = -\sum_i \sum_j C(i,j)\log C(i,j), \quad (5)$$

$$I = \sum_i \sum_j (i-j)^2 C(i,j), \quad (6)$$

$$H = \sum_i \sum_j \frac{1}{1+(i-j)^2} n(i,j), \quad (7)$$

where C is the co-occurrence matrix. Various classifiers including support vector machines (SVM), linear discriminant analysis (LDA) and neural networks have been employed with these features for detection and classification of leather defects.

Pistori *et al.* presented a comparison performed for several classifiers using the Concurrence Matrix's first - order descriptors [12]. The classifiers were applied on both samples of rawhide and wet blue leather. Datasets were trained and tested at different image sizes ranging from 10×10 pixels to 40×40 pixels. In both rawhide and wet blue leather SVM and KNN both correctly classified samples 100% when image size was 40×40 pixels.

Viana *et al.* adopts the Support Vector Machine as a classifier with a focus on properly adjusting the classifier parameters using a stochastic method [13]. They compared LIBSVM and Weka SMO and also used Simulated Annealing for parameter tuning. Ticks, brand marks, cuts and scabies faults were used for detection. Highest accuracy of 99.59% was achieved when LIBSVM was tuned with Simulated Annealing.

Yeh and Perng used image processing technique to identify the defect and group nearby defects into a larger scrap area [14]. They categorized the defects into 7 types which include spots, thin spots, stripes, holes, lines, irregulars, and patterns. They were able to achieve an error rate of 1.16% on lot size greater than 600. In total, 170 samples were used in this research. Jian *et al.* used Feed Forward Neural Network (FNN) analysis for obtaining attributes of leather surface defects and they then applied an algorithm decision tree for classification of defects [15]. They achieved an accuracy of 90%. Villar *et al.* presented an automatic Wet Blue Leather defect classification method [16]. The classifier used was a multilayer perceptron. This method worked by applying the Sequential Forward Selection method in the selection stage of the feature, as well as an appropriate procedure to train the neural network. Their success rate was higher than 95%. Pereira *et al.* used computer vision for goat leather classification. 10 different type of defects were used during training [17]. The classifiers used were KNN, MLP, and SVM. Among all these classifiers SVM with RBF kernel gave the best accuracy rates. Amorim *et al.* used attribute reduction techniques to apply four different classifiers [18]. The attribute reduction techniques used were CLDA, Fisher Face, YLDA, DLDA, and KLDA. They used C4.5, kNN, Naïve Bayes and SVM classifiers. They trained and tested classifiers on both wet-blue and rawhide images. On wet-blue images, kNN performed better with CLDA used to attribute reduction. It was able to achieve a correct classification rate (CCR) of 90.3%. When raw-hide images were used again kNN classifier and CLDA outperformed others and were able to achieve a CCR of 92.23%.

D. MISCELLANEOUS METHODS

The LeaVis system was the first system that used machine vision to detect leather defects [19]. The LeaVis system was designed to process and segment various images of defects and areas of different quality, that are usually manually tagged/marked. Poelzleitner and Niel followed a hierarchical approach and was able to detect seven features of wet-blue leather [20].

Krastev *et al.* showed a histograms-based detection method, using the χ^2 criteria for image analysis and histogram construction [21]. This method detects leather defects based on evaluating the distinction between the grayscale histogram and other image search areas. Yeh and Perng evaluate semi-automatic methods for extraction and classification of wet blue leather defects. Compared to human specialists, their results are reliable and efficient [22]. A fully quantified system for measuring leather raw hides called the demerit count reference standard is the main contribution of the work. The authors also point out that the need for specialized human intervention to take account of the total amount of demerit on wet blue leather is one of the disadvantages of their proposal.

Kwak *et al.* presented a method based on the use of geometric and statistical descriptors in addition to the use of decision trees for the classification of leather surfaces [23]. After Wet Blue, the leather classification is performed on finished leather levels, so a classification mistake is the irrevocable loss of the leather piece. Georgieva proposed the chi-square criteria to analyze the histograms of the image [24]. This method compares the distance with the defects between the histogram for the area analyzed and the area histogram. Amorim *et al.* implemented software to manually extract and label samples from defective areas. Kasi *et al.* proposed an auto-adaptive edge detection algorithm [25]. The algorithm is claimed to outperform the already available edge detectors which include Canny, Sobel, Prewitt and Roberts. This algorithm gives clear and continuous edges as compared to other algorithms.

Ho *et al.* used a multi-camera fusion-based system used for marking leather defects [26]. Multi-camera calibration was achieved by applying homography transformation. Overlapping of pixels was calculated using a homography matrix and boundary re-sampling was then implemented to blend images. The authors claim that their method improved the percentage error obtained by manual processing from 1.1% to 0.1%.

Penaranda *et al.* implemented a computer-vision based artificial system for leather distribution, inspection, and water-jet cutting [27]. Anand et. al used a computer vision system as the front end to obtain the image of each uneven sheet and part and thus solve the problem of two-dimensional stock cutting in the apparel and leather industries [28]. Paakkari et al. used computer vision to provide precise real-time information on the preform outline and position used in the test phase, resulting in cost and material savings [29]. Lanzetta and Tantussi's vision-based leather trimming laboratory prototype was suggested to raise the level of automation in the leather sector [30].

IV. COMPARATIVE ANALYSIS OF IMAGE PROCESSING BASED METHODS FOR LEATHER DEFECT DETECTION

In this section, we compare the performance of methods discussed in the previous section in terms of their leather classification capability. Table 3 presents a comparison of methods in terms of their classification accuracy for leather defect detection. It is noteworthy that Table 3 presented here does not fully represent all the methods discussed in the previous section. The reason for this is the inconsistent performance evaluation of previous methods and lack of a common benchmark. In this Table, we present only those methods which present their standard classification accuracy performance. A majority of the methods in the literature for leather defect classification only report the performance measure of their choice on their data. This is one of the main reasons which causes hindrance in their thorough comparative evaluation. In addition to the lack of a suitable benchmark another problem that hinders the thorough comparative evaluation of leather defect classification methods is the lack of publicly available software/code against the reported methods.

It can be seen from Table 3 that most methods report above 90% accuracy, while the KNN approach achieved 100% accuracy. This performance can be partially attributed to the fact that all these methods are evaluated on very small local datasets. The largest collection of leather images used in the aforementioned studies consists of 700 images out of which 70% were used for training and 30% for testing. Hence, the largest test set used for evaluation by these studies is comprised of only around 200 images. Considering the natural variations that can occur in leather samples in industrial processing, this seems to be a small test set for evaluation of models.

Secondly, the top performing methods such as the method of Pistori et. al [12] evaluated their model on only 16 images, while Viana *et al.* [13] evaluated their method on only 15 images. They extract train and test patches from the small number of leather images to demonstrate their efficiency. Although these studies employed powerful classifiers (such as SVM with quadratic kernel as in Viana *et al.* [13] and k-nearest neighbor classifier as in Pistori *et al.* [12]) in addition to robust feature attributes based on texture and color information, the generalization of these methods on such small datasets may not guarantee robustness on highly varying real-world leather imagery. The outstanding performance of these techniques can also be attributed to their high dimensional feature sets such as the 139-dimensional attribute set in the work of Viana *et al.* [13] and a 66-dimensional feature set employed by the technique of Pistori *et al.* [12]. The technique of Jawahar *et al.* [10] employed high dimensional Wavelet based statistical and co-occurrence features to tackle the problem of leather defect classification. The high dimensional representation in the feature space enabled the SVM classifier with a Gaussian kernel to achieve a test accuracy of 98.8% on 200 leather images. It would be interesting to evaluate the approach of Jawahar et. al on real-world varying leather samples in industrial settings for categorization of multiple defect types.

Table 4 compares image-analysis based leather defect detection methods in terms of their defect detection capability. In other words, it presents the maximum number of defects that each method is capable of predicting. Classifier based methods including SVM and decision trees can detect

TABLE 3. Comparison of leather defect detection methods in terms of classification accuracy .

Sr. #	Authors	Model	Code and Models	Brief Description	Classification Accuracy (%)
1	R. Viana et. al [13]	LIBSVM	not available	A machine learning classifier is trained to learn the mapping function between a set of features and leather defect type.	99.6
2	M. Jawahar et. al [10]	WSF+WCF	not available	Binary SVM with Gaussian kernel was used to discriminate defective and non-defective leather samples.	98.8
3	C. Kwak et. al [23]	Decision tree	not available	Used sequential decision-tree classification scheme in order to maximize the classification efficiency.	91.3
4	L. Jian et. al [15]	FFN decision tree	not available	FNN is used for selecting the various relevant attributes, to get the properties for the best combination of classification, and then we can extract classification rules by the decision tree method.	94.0
5	M. Jawahar et. al [32]	ANN	not available	Used Artificial Neural Network for defect detection due to their ability to describe complex decision regions, and Artificial Neural Network is one of the most flexible classifiers.	88.6
6	P. Villar et. al [16]	NN	not available	Bayesian Regularization algorithm is used for training a neural network because it offers better training speed and a method to determine the number of neurons in the hidden layer.	96.6
7	R.F. Periera et. al [17]	MLP	not available	Used multilayer perceptron. In the first step it finds the failure regions and in next the one extracts some features from the failure map it found.	90.3
8	H. Pistori et. al [12]	KNN	not available	The k-nearest neighbor approach was used using Weka's implementation. Tested Knn with 5 neighbours, weighted by the inverse of their distance.	100.0

up to four defects, whereas image-analysis based methods improve the number of defects to greater than five. Notably, ANN and KNN based methods outperform other classifier and image analysis based methods by detecting more than 10 different leather defects.

V. PROPOSAL FOR APPLIED DEEP LEARNING FOR AUTONOMOUS LEATHER QUALITY INSPECTION

Based on our proposed guidelines for autonomous leather quality grading in Section I, in this section, we provide a brief background knowledge on deep learning and specifically convolutional neural networks in the context of visual leather inspection. Next, based on prior works and the requirements of the leather inspection process, we provide recommendations and guidelines for choosing appropriate architectures for leather defect detection and categorization, semantic segmentation of defects for in-depth analysis and multi-class leather grade classification.

A. DEEP LEARNING BACKGROUND

In the leather industry, usually two types of machine learning are required: defect detection (classification problem), and in-depth geometric information of the detected defects (segmentation problem). Leather images have so many variations in terms of morphology and contrast of defects. There could be more than 15 defects in one image with

TABLE 4. Comparison of methods in terms of their defect detection capability .

Sr. #	Authors	Model	Binary Classification	4 defects	≥ 5 defects	≥ 10 defects
1	R. Viana et. al [13]	LIBSVM		✓		
2	M. Jawahar et. al [10]	WSF+WCF	✓			
3	C. Kwak et. al [23]	Decision tree			✓	
4	L. Jian et. al [15]	FFN decision tree		✓		
5	M. Jawahar et. al [32]	ANN				✓
6	P. Villar et. al [16]	NN		✓		
7	R.F. Periera et. al [17]	MLP	✓			
8	H. Pistori et. al [12]	KNN				✓
9	H. Pistori et. al [18]	Attribute reduction			✓	
10	C. Yeh et. al [22]	Custom			✓	
11	C. Yeh et. al [14]	Digital Image Processing			✓	

different contrasts. Even one type of defect itself varies greatly in an image. This is the reason why traditional machine learning algorithms fail to detect multiple defects at one time. This makes the leather defect classification/ segmentation problem the best candidate for deep neural networks. The good thing about neural networks is that they can be designed for both image classification and image segmentation. Classification is the process of taking an input (i.e. image) and outputting a class or probability that the input is a particular class. Image segmentation is the task of partitioning the image in to multiple segments. Semantic segmentation is used to do this in deep neural networks. Semantic segmentation is the task of assigning a class to each pixel in a given image. Classification assigns a single class to the whole image, while semantic segmentation classifies each pixel of the image to one of the classes. Some essential processes common to both applications are

- transfer learning,
- data augmentation,
- loss function.

CNN's are currently one of the most widely used architectures for deep image learning. Data-related features need to be extracted manually in traditional machine learning, while deep learning employ the raw image for learning features. CNN's structure consists of a layer of input, a layer of output and several hidden layers sandwiched between layers of input and output. The hidden layers consist of convolutional layers, max-pooling layers, and layers that are fully connected. The general architecture is shown in Figure 9.

CNN architectures vary in the type and number of layers implemented for any specific application. If we have a continuous application then at the end the CNN needs to have a regression layer. If the application is for a categorical response then the last layer must be a classification layer. Neurons are in 3D shape in CNN layers. The layers hold the height, width, and color dimension RGB.

There are several models available for classification and semantic segmentation. The model structure shall be chosen properly depending on the use case. There are several things which should be taken into account. In the case of leather, these are the size of an image, number of training images, number of defects and the variation in defects. Usually, deep learning-based segmentation models are built upon a base CNN network. Some initial layers of the base network are used in the encoder, and rest of the segmentation network is built on top of that. For most of the segmentation models, any base network can be used.

For classification and semantic segmentation, our first task is to select an appropriate network. One of the advantages of segmentation is that the classification models can be used for segmentation we just have to add a decoder in classification algorithms. A standard model such as ResNet [33] or VGG [34] is chosen for the base network usually. ResNet is the model proposed by Microsoft which got 96.4% accuracy in the ImageNet [35] 2016 competition. ResNet is used as a pre-trained model for several applications. ResNet has a large number of layers along with residual connections which make its training feasible. VGG-16 is the model proposed by Oxford which got 92.7% accuracy in the ImageNet 2013 competition. Compared to Resnet it has lesser layers, hence it is much faster to train. For most of the existing segmentation benchmarks, VGG does not perform as well as ResNet in terms of accuracy. GoogLeNet aka inception v3 was ILSVRC-2014 winner [36]. It is a 22-layer model

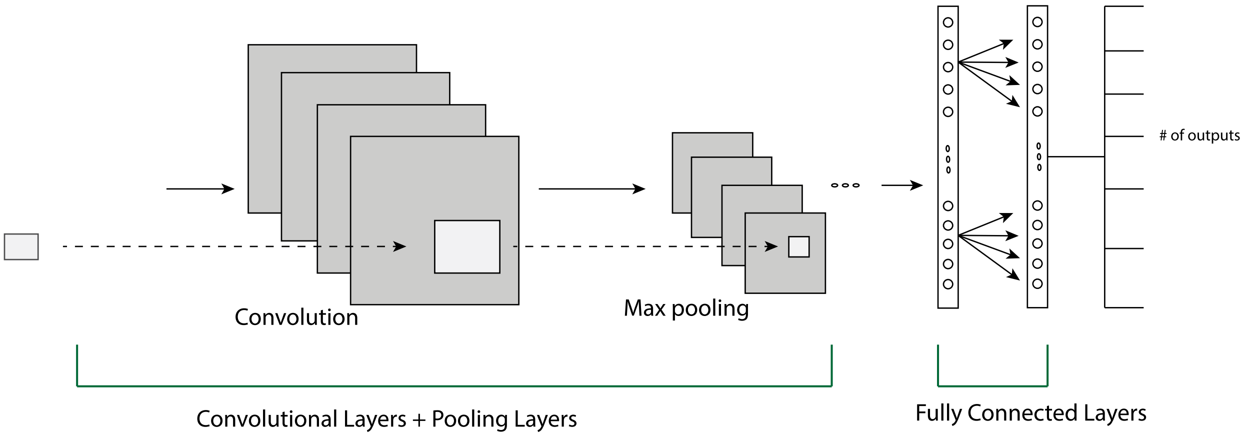

FIGURE 9. The general architecture of a convolutional neural network. The image is first subjected to a set of 3-dimensional filters by the convolutional layers. Subsequently, information is down-sampled through max-pooling layers for effective classification. Finally, a set of fully-connected layers map the 3-dimensional volume to output nodes for classification.

with an error rate of 6.7%. The main contribution of this architecture was that the addition of Inception Module which dramatically reduced the number of parameters. Since some algorithms are based on contrast some on edges, therefore, choosing an appropriate algorithm for leather application (with varying contrast and morphology) is a researchable problem. There are also limitations on the size of these algorithms. Most algorithms give good accuracy to small image sizes, which may not be suitable for multiple defect detection in leather images where the morphology of the defect varies greatly. Apart from these models, we can also train a custom network. One of the advantages of using a custom model is that we can customize it according to our application that is new to deep neural networks. Some algorithms work for many applications in which feature morphology is almost similar among those applications. For such applications, choosing a model pre-trained on ImageNet is the best choice. For a pre-trained model, transfer learning is used.

Deep network learning from scratch is typically not feasible because there are several reasons: a data set of sufficient size is required (usually not available in case of leather) and convergence may take too long to be worth the experiments. In case of the small dataset, starting with pre-trained weights is often helpful rather than randomly initialized weights. One of the major transfer learning scenarios is to fine-tune the weights of a pre-trained network by continuing with the training process. However, it is not entirely straightforward to apply this transfer learning technique. To use a pre-trained network, there are architectural limitations that must be met. In the case of leather, there are not many pre-trained models available. Most closely related are fabric defects. We can use the weights of these models for transfer learning. The dataset for wet blue leather is not available online. Researchers working on leather defects usually have small datasets of 60-80 images. These images are not enough to train a deep learning model, and not best suited for learning weight from scratch. One reason not to have a big dataset for leathers is that most of the industry is reluctant to share its data with the researchers. On the other hand, this field is also ignored by the researchers. For example, in medical fields, there are many competitions held annually that release large datasets to work on a certain problem. But in leather, there are no such competitions. In leather to get a large dataset, there is only one option that is data augmentation.

Data augmentation is a popular method that has been demonstrated to aid machine learning models in general and deep learning architectures specifically; either acting as a regularizer or to speed-up convergence, thereby avoiding overfitting and improving generalization capabilities. The objective of these transformations is to produce more samples to create a bigger dataset, avoid overfitting, model regularization, keep the equilibrium of each class in the dataset, and even synthetically generate new samples that are more representative for the use case or task. Data augmentation will not only increase the number of images in our dataset but it will also add variations in the image. Some images will be zoomed in, zoomed out, have different contrast, etc. So, the dataset will have more images and the algorithm trained will also be able to detect most of the images under different lighting conditions. After data augmentation, we will get a large dataset with multiple defects on each image.

Selection of the right algorithm is not the only solution. There are many other parameters that we need to take care of as well. Hyper-parameters is one of them. In hyper-parameters we have learning rate. We need to choose a suitable learning rate policy, usually the lower rate is preferred as the pre-trained weights are comparatively good and there is no need for drastic change. In custom algorithms, learning rate selection is a very complex job. Because setting a learning rate depends on the size of the image, size of the dataset and hardware (GPU). To overcome this problem we can use batch normalization. Batch normalization allows each layer of a network to learn by itself a little

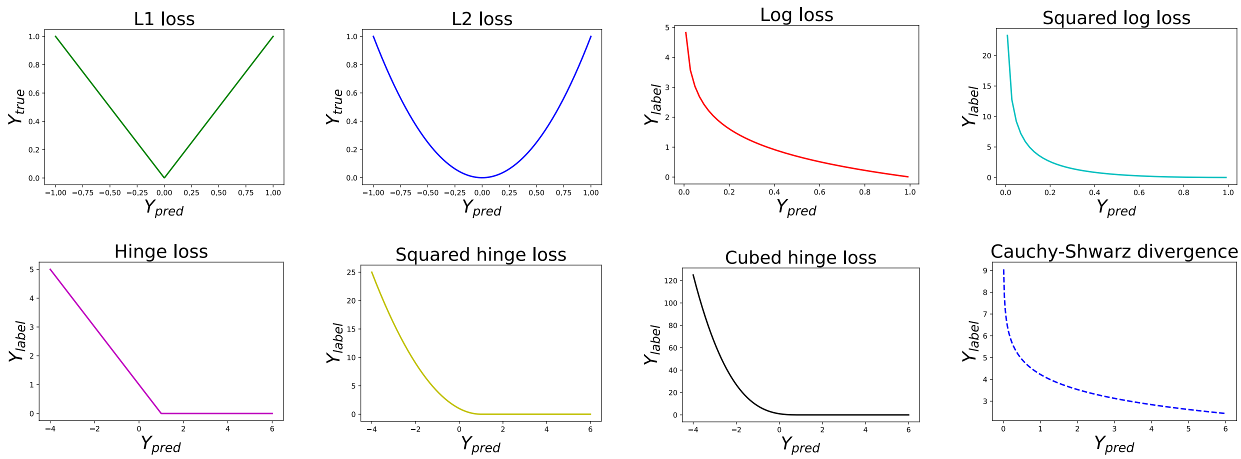

FIGURE 10. Common loss functions used in deep learning.

bit more independently from the other layers. We can use higher learning rates because batch normalization makes sure that no activations gone high or low. And by that, things that previously could not be trained, will start to be trained. It reduces overfitting because it also has slight regularization effects. Therefore, if we use batch normalization we will use less dropout, which is a good thing because we won't lose a lot of information. Global Average Pooling before dense layer with batch normalization is another good technique that is used for this purpose. The global average pooling mechanism computes the mean value for each feature map and supplies it to a softmax layer. The softmax layer takes each value and converts it to a probability (with the probability of all digits summing to 1.0). It removes the flatten layer which speeds up the training of the model. The elimination of all these trainable parameters also reduces the tendency for overfitting, which needs to be managed in fully connected layers by the use of dropout.

Loss functions are an important component of any deep neural net as they evaluate the predictions of the algorithm on the given data. Together with optimization techniques, the loss functions enable the net to learn to reduce the prediction error gradually. Loss functions are generally classified into regression and classification losses given the task at hand. A few common loss functions are presented in Table 5 and are also plotted in Figure 10.

L_1 and L_2 losses presented here generally belong to the class of regression losses. They measure the average sum of absolute differences or the average of squared differences between the gold standard and predictions. The most common classification loss functions can be categorized either as hinge loss (also known as SVM loss) or log loss (also known as cross-entropy loss). Hinge loss advocates that the score of the correct category should always be greater than the sum of all incorrect category scores by some safety margin. Therefore, it is most notably used for maximum-margin classification mainly in an SVM framework. Cross-entropy loss provides a direct relationship between the predicted probability and the actual label, increasing as the predicted probability diverges from the actual label. An important aspect of this is that the cross-entropy loss heavily penalizes confident but wrong predictions.

In the case of leather, there are many loss functions which can be used depending on the number of defects and size of defects. Tang et al. showed that well-fitted hinge loss can outperform log loss based networks in typical classification tasks. Comparing various loss functions under different conditions and provide insights into when and which function can be used for leather defects is an interesting avenue to explore. Janocha et. al showed that on the MNIST dataset without any data augmentation L1 performed better in accuracy [37]. In case of loss function which converges faster $hinge^2$ and $hinge^3$ losses are the fastest in training, and once the number of hidden layers is increased then L2 also gives better results.

Janocha et. al also showed that Cauchy-Schwarz Divergence as the optimization criterion seems to be a consistently better choice than log loss [37]. It performs equally well or better on both MNIST and CIFAR10 in terms of both learning speed and the final performance. Table 5 presents a variety of loss functions that have proven useful in different settings. But which loss function performs well on leather is an open-ended problem. Due to the large variation in morphology and contrast, it may be necessary to look for a custom loss function that performs well on leather related problems.

B. RECOMMENDATIONS FOR CHOOSING CNN BASED ARCHITECTURES FOR LEATHER DEFECT DETECTION AND CATEGORIZATION

Numerous deep learning-based object detection frameworks have been proposed to date. These frameworks can be generally classified into region proposals based methods and single-shot based detectors. Region proposals based object detection frameworks have three stages: region proposal

TABLE 5. List of loss functions employed in deep networks for classification .

Name	Equation	Symbol
L_1 loss	$\|\mathbf{y}-\mathbf{p}\|_1$	$\mathcal{L}_1$
L_2 loss	$\|\mathbf{y}-\mathbf{p}\|_2^2$	$\mathcal{L}_2$
log (cross entropy) loss	$-\sum_j \mathbf{y}^{(j)} \log \mathrm{E}(\mathbf{p}^{(j)})$	log
squared log loss	$-\sum_j [\mathbf{y}^{(j)} \log \mathrm{E}(\mathbf{p}^{(j)})]^2$	$\log^2$
hinge loss [38]	$-\sum_j \max(0, \frac{1}{2} - \hat{\mathbf{y}}^{(j)}\mathbf{p}^{(j)})$	hinge
squared hinge loss	$-\sum_j \max(0, \frac{1}{2} - \hat{\mathbf{y}}^{(j)}\mathbf{p}^{(j)})^2$	hinge^2
cubed hinge loss	$-\sum_j \max(0, \frac{1}{2} - \hat{\mathbf{y}}^{(j)}\mathbf{p}^{(j)})^3$	hinge^3
Cauchy-Schwarz divergence [39]	$-\log \frac{\sum_j \mathrm{E}(\mathbf{p}^{(j)})\hat{\mathbf{y}}^{(j)}}{\|\mathrm{E}(\mathbf{p})\|_2 \|\hat{\mathbf{y}}\|_2}$	$\mathrm{D_{CS}}$

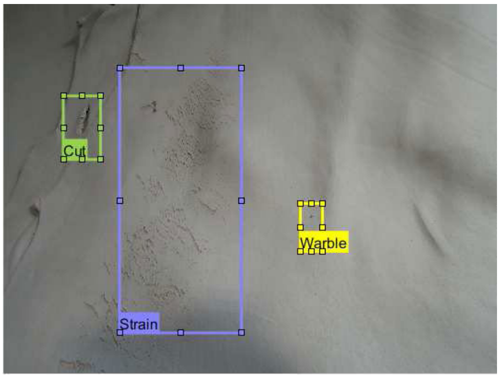

FIGURE 11. Desired output of the defect detection and classification stage. The strain highlighted by a bluish bounding box is a mechanical defect. The warble defect highlighted by a yellow bounding box is a minor defect and generally ignored specifically in New Zealand leather.

generation module, a convolutional neural network as a feature extractor and a classification plus regression module for generating multiple classes and bounding boxes. A few popular region proposal based methods which are discussed here for application to leather defect detection include R-CNN [40], Fast-RCNN [41] and Faster-RCNN [42]. In contrast to region proposal based networks, single-shot based detectors are one-step and directly map image regions to bounding boxes and corresponding class probabilities based on global classification/regression, thus having reduced training and inference times. A few single shot detectors ranging from pioneers to recent ones are discussed here according to a leather defect detection perspective. The problem of leather defect detection and classification is depicted with the help of a representative example in Figure 11. For a leather sample, we are interested to effectively localize all leather defects and recognize their type simultaneously. The desired output of our defect detection framework is shown in Figure 11 for a representative leather sample that includes three categories of defects, namely; cut, shear and wart.

Important parameters that must be considered for the leather defect detection and classification task are *image resolution, the scale of defects and the variations in defects*. Leather images are usually high in resolution and resizing them for computational efficiency is not always viable as some of the defects occur at a tiny scale. The variation of defects in terms of shape, texture, color, and contrast is another important aspect which requires rich feature representations to cater for all variations. Therefore, the architectures for leather defect detection must be designed carefully in light of these important parameters.

Region proposal based networks may not be able to tackle high-resolution leather images due to their considerable latencies at training and test stages. These networks also may not be robust against defects of various shapes and scales, due to their object-oriented region proposal computation strategy. The baseline object detector R-CNN proposed by Girshick *et al.* was demonstrated to capture rich feature representations for a wide variety of object categories, however, the following reasons limit its adaptability to the leather defect detection and classification problem [40]. R-CNN either warps or crop region proposals to make the objects of interest (i.e., leather defects in our case) equally sized [40]. These operations might cause geometric distortions of the leather defect or part of the defect to be cropped, thus affecting detection and classification accuracies. Also, the selective search methods for object proposals employed in R-CNN promotes only larger candidate regions, which is not suitable to detect leather defects that are at a very small scale [43]. Due to the parameter heavy CNN used by the R-CNN (i.e., VGG16 [34]), it takes 47s/image on a GPU at the test stage. Therefore, it is not a viable option for high-resolution leather images. Fast R-CNN is better than R-CNN in terms of the reduced train and test times, which is achieved through feature sharing during training [43]. However, the region proposal pipeline of Fast R-CNN is similar to R-CNN, which limits its application in the leather defect detection task. The Faster R-CNN improves upon the region proposal problem of R-CNN and Fast R-CNN by introducing a fully convolutional Region Proposal Network (RPN) [42]. The RPN produces better proposals as compared with the former region proposal based methods, however, it is not robust against extreme scales or shapes. To customize region proposal based methods for leather defect classification, the region proposal framework must be adapted as per the scale and shape of the leather defects. The region proposal based architectures may also require optimization to tackle high-resolution images.

Single-shot based detectors can tackle higher resolution images when compared to the region proposal based networks, owing to their reduced pipeline. They might also be able to better handle the scale and shape variations. The pioneering work You Only Look Once (YOLO) divides the image into an S $\times$ S grid, where each grid cell predicts bounding boxes and corresponding confidence scores [44]. This scheme may allow YOLO to detect leather defects at multiple scales depending upon the grid size S. However, YOLO is not robust at detecting groups of small objects such as pinholes or a group of warts and also fails to generalize objects with unusual aspect ratios, which is a common feature of leather defects. The Single Shot MultiBox Detector (SSD) on the other hand can handle objects with various sizes (defects in our case) by using anchor boxes with various aspect ratios and fusing predictions from feature maps of various resolutions [45]. The single-shot detector may also be tuned to detect small defects by adopting a better feature extractor, adding context through deconvolution layers with skip connections and better network structure [46]. Finally, RetinaNet is also worth investigating for leather defect detection due to its ability to handle higher resolution images and detect defects at multiple scales and shapes as compared with the rest of the detectors. The mentioned capabilities of the RetinaNet may be attributed solely to its feature pyramid network [47].

C. RECOMMENDATIONS FOR DEEP LEARNING BASED SEMANTIC SEGMENTATION FOR LEATHER DEFECT ANALYSIS

For grading hides, more information about defects is required as sufficed by the detection and classification pipeline. This information includes various modalities such as the area of the defect, total area of the hide covered by the defects and distance of the defects from the perimeter of the skin. For the precise computation of this information, we need accurate semantic segmentation of all defects to pixel-level accuracy. The desired output of the segmentation and defect analysis module is depicted by a representative example in Figure 12, which can be effectively utilized for leather quality grading.

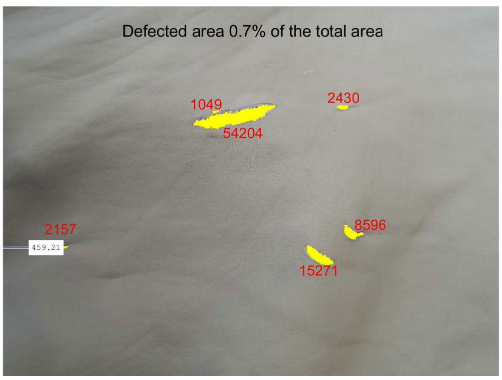

FIGURE 12. Desired output of the defect analysis module with relevant area and distance measurements.

Fully convolutional networks (FCNs) have recently exhibited excellent progress in semantic segmentation in applications ranging from real-world scenes, medical image segmentation and industrial defect inspection. Here, we only review representative FCNs which may be suited to leather defect segmentation. A few important considerations for designing FCNs for leather defect segmentation include the ability of the architecture to tackle high-resolution leather images, ability to handle defects at multiple scales including tiny defects and an ability to capture accurate shape information of the defects.

1) HIGH-RESOLUTION LEATHER IMAGES

for most FCNs designed for semantic segmentation such as the work of Long *et al.* [48] or encoder-decoder architectures such as SegNet [49], the memory requirements of their backbone architectures determine the image resolution that they can support. There exists a trade-off between accuracy and resource requirements. For instance, ResNet-152, when employed as the backbone, may result in improved accuracy as compared to ResNet-50 [33]. However, this improved accuracy will be achieved at the cost of ∼2 times the memory requirements, which can further scale up with the resolution of the leather images. On the contrary, the U-Net architecture proposed by Ronneberger *et al.* [50] for high-resolution medical image segmentation is a good contender for leather defect segmentation. Its tiling strategy and a relatively compact architecture make it suitable for high-resolution images while keeping the memory requirements at a low end.

The more recent methods such as the DeepLab based networks [51]–[53] are memory heavy as they employ the likes of VGG16 [34] or ResNets as their backbone architectures. However, they can still cope with high-resolution images and therefore can be adapted to the task of leather defect segmentation.

2) MULTISCALE LEATHER DEFECTS AND SHAPE INFORMATION

are other important considerations for leather defect segmentation. For leather segmentation, SegNet can be adapted to preserve shape information through skip connections for robust segmentation, however, its ability to handle defects at multiple scales must be improved for practical use [49]. Contrarily, U-Net can be adapted to handle defects at multiple scales and can be modified to preserve shape information as well as [50]. The DeepLab based architectures [51] are specialized in capturing multiscale information and preserving shape information. Specifically the work of Chen *et al.* [54] (based on DeepLab) can be investigated for boundary/shape preservation of defects and the recent work of Chen *et al.* [55] can be adapted to obtain scale-aware pixel-level accurate segmentation of leather defects.

D. GUIDELINES FOR DEEP NEURAL NETWORKS BASED MULTICLASS LEATHER GRADE CLASSIFICATION

According to Section I, multiclass classification is required for leather quality grading given the feature matrix from the image analysis module. This implies that the accuracy of the grading stage depends upon the information extracted about the defects in the detection, segmentation and image analysis stages. Consequently, the design and complexity of the multiclass classification stage depends upon the robustness of the previous steps in the pipeline. A wide variety of deep learning architectures exist in literature that have been successfully applied in similar problems and can be easily adapted for the task at hand. However, it is vital to mention here that the multiclass classification based deep network for leather

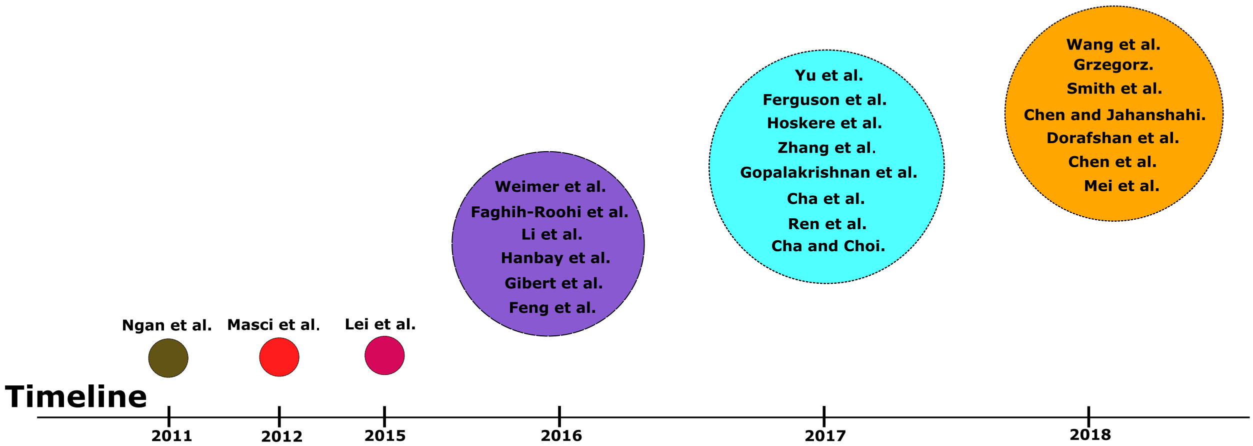

FIGURE 13. A timeline of recent progress in deep learning for visual defect inspection. The cited methods depict representative methods on the subject. It can be observed that there is a growing trend in deep learning-based solutions for visual defect inspection. The methods published in recent two years (2017,2018) are greater than double the number of works published till 2016. Based on this healthy trend, more deep learning-based automated inspection systems are expected for the surveillance industry in the near future.

quality grading must be able to effectively utilize information from multiple expert human graders, which might enable it to match or surpass human level performance. This is a researchable problem and needs investigation. The rest of this subsection is dedicated to reviewing recent CNN-based models that are employed for multiclass classification in visual defect screening applications. The success of these methods in highly related tasks may promote their potential application for multiclass classification based leather quality grading.

In Figure 13 and Table 6, we compile recent methods that employ deep neural networks for visual inspection in related applications for both defect detection and multiclass classification. Here, we are interested in methods proposed for multiclass classification of defects, therefore, we only provide details of those models in this section. It can be observed that there is an increasing trend in the number of methods that employ CNNs for automated inspection in various applications. The increasing number of publications in 2016-2018 is a clear indication of the success of CNN based methods in machine vision applications for automated inspection.

Dorafshan *et al.* performed transfer learning on concrete images using AlexNet [72] for crack inspection [56]. Their method comprehensively outperformed traditional edge detectors on cracks of all types. Gopalkrishnan *et al.* employed a truncated version of the VGG16 architecture and performed transfer learning on pavement images for crack detection [57]. The features generated by the CNN were classified using various machine learning classifiers, where the best performance was achieved by a single layer neural network classifier (VGG-16 DCNN). Chen and Jahanshahi dealt with the very important subject of crack detection on nuclear reactor surfaces [58]. They proposed a CNN based architecture with a Naïve Bayes based data fusion scheme. Additionally, a novel data fusion scheme was proposed for multi-frame data aggregation from nuclear inspection videos.

Ferguson *et al.* examined how to use various CNN architectures to localize casting defects in X-ray images [62]. The authors employed a defect classification model on a sequence of defective pictures to categorize various types of defects and then used a sliding window method to develop a localization model. Weimer *et al.* investigated a new paradigm that examined Convolutional Neural Networks (CNN) design configurations and the impact of various hyperparameter settings on the accuracy of the results of defect detection [63]. Psuj proposed a CNN based method for defect categorization in steel elements [66]. Specifically, rectangular-shaped artificial defects were employed to evaluate the efficacy on the method of steel elements. Masci *et al.* employed a Max-Pooling CNN approach for supervised steel defect classification tasks with 7 defects [68]. A 7% error rate was obtained.

Ren *et al.* proposed a general approach requiring a small amount of training data for automated surface inspection [69]. Transfer learning is employed on a pre-trained DeCaf model to generate features which are classified by a multinomial logistic regression classifier for defect categorization. Feng *et al.* proposed a binary defect identification (injurious or non-injurious) method from magnetic flux leakage (MFL) images based on a CNN [70]. Unlike prior approaches, instead of the MFL measurement features, this technique is fed by the MFL images. To demonstrate the efficiency of the suggested model, the authors use real MFL data gathered from experimental pipelines.

VI. DATA

In the literature, many studies evaluate the detection of leather defects. It should be noted that different studies use contrasting datasets and possibly distinct parameter settings. Nelore and Hereford cattle (Brazil) is a dataset with 50 images of wet blue leather pieces. Amorim *et al.* used this dataset in their

TABLE 6. CNN based methods for defect inspection targeting different applications. The training information along with architectural information is provided. Here 'S' is the number of training images; column represents subimages, where 'P' stands for image patches. The details of the baseline architecture employed are also presented. The techniques presented in this Table can be generally categorized into the following applications: crack detection, fabric defect detection, industrial defect detection, and steel defect detection. Also, each method can be categorized either as a fully convolutional (Fully Conv) network or a classical convolutional based network.

Sr. #	Model	Pub	Year	Number of Training Images	Pre-trained Model	Fully Conv	Application
1	DCNN, [56]	CBM	2018	12809 (S)	AlexNet	✗	Crack Detection
2	DCNN_TL, [57]	CBM	2017	760	VGG16	✗	
3	NB-CNN, [58]	TIE	2018	5326 (P)	—	✗	
4	MSCDAE, [59]	Sensors	2018	—	—	✓	Fabric Defect Detection
5	FCSDA, [60]	TASE	2016	2600 (P)	—	✓	
6	SSD, [61]	TIM	2018	6371	VGG16, YOLO	✓	Industrial Defect Detection
7	XnetV2, [62]	ICBD	2017	23760 (P)	ResNet-101 [33]	✗	
8	DII, [63]	CIRP Annals	2016	1,299,200	—	✗	
9	FCN-SDI, [64]	ICCVS	2017	720	R-CNN	✓	
10	DCNN-L, [65]	IJCNN	2016	22408 (S)	—	✗	Steel Defect Detection
11	DMTL, [66]	TITS	2016	62500 (S)	—	✓	
12	DCNN-DND, [67]	Sensors	2018	—	—	✗	
13	MPCNN, [68]	IJCNN	2012	2281	—	✗	
14	Decaf+MLR, [69]	TC	2017	—	DeCAF	✗	Miscellenous
15	CNNMFL, [70]	TIM	2016	25650	—	✗	
16	DFRCNN, [71]	CACIE	2017	2366	Fast R-CNN	✓	

research [18]. There is another dataset from the DTCOURO project [73]. This dataset is part of a scientific research and technological development project, DTCOURO, that envisages the development of a computer-based, fully automated system for the classification and grading of rawhide and leather in bovine animals. One of the objectives of the DTCOURO project is to propose and provide comparative studies of pre-processing, feature extraction, feature selection, segmentation, and classification techniques. Table 7 presents the details of the various datasets employed by previous methods for leather defect detection. It can be observed that apart from the DTCOURO dataset, all other datasets are relatively small thus limiting extensive evaluation of the developed algorithms. Also, despite developments in the field, unavailability of datasets is a major hindrance in progress in this field. As presented in Table 7, none of the studies make their data publicly available for comparative evaluation of algorithms and benchmarking. This urges a need for the development of a sufficiently large publicly available dataset for comparative evaluation and to test the generalization ability of the methods. To address this issue, we are in the process of constructing a large database of wet blue leather images for defect classification and quality grading. The dataset will be publicly available for benchmarking and comparative evaluation of algorithms.

Data collection is the most significant component of this study, as there is no dataset for leather defects available. Liong *et al.* used a six-axis desktop robotic arm, a high-resolution camera, 3d printed plastic parts and a non-flickering LED light source [74]. Bong *et al.* set up an image-grabbing scheme [75]. The grabbing system consists of a bracket used to hold a camera and a light source. The fabric is spread over the table under the bracket. The camera is combined with a finite aperture lens for focusing the image. Unfortunately, in the real workspace, the corresponding point in the image and the optical center of the camera are not collinear. These are some of the techniques that people usually use. Almost everyone used an HD camera to collect images. In the leather industry, the material continues to move and needs to be inspected for defects. In situations where ongoing material needs to be checked for faults, line-scan cameras will usually provide a better alternative than traditional techniques. Despite this, people frequently harbor reservations about deploying line-scan cameras; reservations that usually emerge from the inadequate experience with this technique of inspection. This often leads to circumstances in which users of image processing often tend to use familiar area scan technology even when the use of line scan cameras makes more sense for a certain application. But the fact is that line scan cameras provide a cost-effective way to generate high-resolution images and to make them available on generic

TABLE 7. Details of the datasets employed by previous studies for leather defect detection.

Sr. #	Authors	Model	Dataset(s) Employed	Number of Images	Publicly Available
1	R. Viana et. al [13]	LIBSVM	DTCOURO	14722	✗
2	M. Jawahar et. al [10]	WSF+WCF	Custom	700	✗
3	C. Kwak et. al [23]	Decision tree	Custom	140	✗
4	L. Jian et. al [15]	FFN decision tree	Custom	200	✗
5	M. Jawahar et. al [32]	ANN	Custom	90	✗
6	P. Villar et. al [16]	NN	Custom	1769	✗
7	R.F. Periera et. al [17]	MLP	Custom	1874	✗
8	H. Pistori et. al [12]	KNN	DTCOURO	2000	✗
9	H. Pistori et. al [18]	Attribute reduction	Nelore and Hereford Tannery, Brazil	2000	✗
10	C. Yeh et. al [22]	Custom	Various tanneries of Taiwan	170	✗
11	C. Yeh et. al [14]	Digital Image Processing	Custom	178	✗

PC platforms for evaluation software whose performance continues to grow.

VII. EVALUATION MEASURES

Many performance measures are widely used in the fields of machine learning, knowledge discovery and data mining. They are primarily used for two purposes. First, they are used as criteria to compare and evaluate machine learning algorithms. Traditionally, a range of measures, such as accuracy, precision, recall, F-measure, and so on are adopted as criteria to evaluate the performance of information systems. For example, in classification tasks, accuracy is defined as the percentage of objects that are correctly classified. It measures the classification performance of a learning algorithm. In the last two decades, accuracy was the most frequently used measure in algorithm performance evaluation. In information retrieval, precision and recall are two traditional measures that are used to evaluate the query quality. In recent years, AUC has emerged as another popular measure in machine learning.

A. CONFUSION MATRIX

A confusion matrix is a table that shows the agreement of the actual values and predicted values for a classification method in a numerical format. As it quantifies the errors obtained by the predictions of classification methods, it is also referred to as an error matrix. A pictorial representation of a confusion matrix is presented in Figure 14. Four key components of the confusion matrix with reference to leather defect classification are defined as:

- **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.

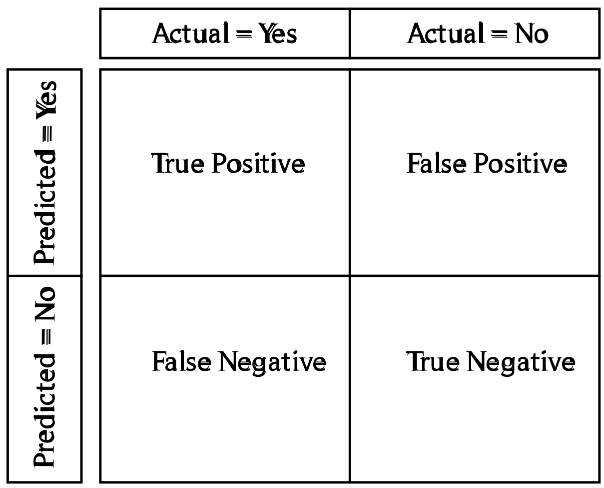

FIGURE 14. Confusion matrix.

- **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.

B. ACCURACY

Accuracy is a statistical measure of how often the prediction made by the classifier correctly identifies a condition. It is the ratio of the number of correct predictions to the total number of predictions, given as:

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

C. LOG LOSS, LOGARITHMIC LOSS OR CROSS ENTROPY LOSS

Log loss, or logarithmic loss, provides more specific details to the classifier. Generally, if the output is a numerical probability rather than a class label, then it is possible to use a log or logarithmic loss. Log-loss is a "soft" accuracy measurement that embodies the idea of probabilistic trust. Mathematical equation of log-loss for a binary classifier is given in equation 9

$$\mathcal{L}_{\log} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1-y_i)\log(1-p_i), \tag{9}$$

where $\mathcal{L}_{\log}$ is the log-loss, p_i is the likelihood that the i-th observation is of class c and y_i is the label in question that is either 0 or 1. Since y_i is either 0 or 1, either the right or the left part to the addition is selected for an instance. Log-loss can be defined as the cross-entropy between the distribution of the true labels and the predictions, and it is strongly linked to what is known as relative entropy. Entropy measures something's unpredictability. Cross-entropy includes true distribution entropy and extra unpredictability when assuming a distribution different from the true distribution. So, log-loss is an information-theoretical metric for measuring "extra noise" that is added due to a predictor as opposed to the true labels. We maximize the classifier's accuracy by minimizing cross-entropy.

D. RECEIVER OPERATING CHARACTERISTICS (ROC) AND AREA UNDER THE ROC CURVE (AUC)

ROC curve is a graph that shows the performance of a classification model at various classification thresholds. It is essentially a plot between the true positive rate (TPR) and false-positive rate (FPR).

- **True Positive Rate(TPR)** is the proportion of the positive data points correctly predicted to all the actual positive data points. It is also referred to as recall. Higher TPR means we will have less positive data points that will be misclassified.
- **False Positive Rate(FPR)** is the proportion of the negative data points falsely predicted as positive to all the actual negative data points. Higher FPR implies that more negative data points have been classified incorrectly.

AUC is the area below the ROC curve. The ROC curve shows the classifier's sensitivity by plotting the rate of actual positive to the false positive. In other words, as we allow more and more false positives to be obtained, it shows how many true positives can be obtained. The perfect classifier that does not make any mistakes would hit a true positive rate of 100 percent immediately without incurring any false positives - this practically never happens.

Although ROC curve provides detailed information about the classifier's behavior, but it is difficult to quickly compare multiple ROC curves. In particular, if some sort of automatic hyper-parameter tuning mechanism were used, instead of a graph requiring visual inspection, the machine would need a quantifiable score. The AUC is one way of summarizing the ROC curve into a single number, making the comparisons of multiple ROC curves simple. Higher values of AUC are better.

E. PRECISION-RECALL

Precision and recall are two distinct metrics but often they are used together. Precision answers the question, "Out of the items that the ranker/classifier predicted to be relevant, how many are truly relevant?" While, recall answers the question, "Out of all the items that are truly relevant, how many are found by the ranker/classifier?"

Mathematically, precision and recall are defined as

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{TP+FN}. \tag{10}$$

VIII. DISCUSSIONS

So far, we have discussed the image analysis based solutions for leather defect inspection, their advantages and shortcomings and the need for deep learning methods for generic visual defect inspection. In this section, we discuss the various challenges that exist in design and deployment of convolutional neural network (CNN) based solutions for automated leather defect inspection. Furthermore, we shed some light on how these challenges can be transformed into opportunities, leading to future research directions in this field.

A. CHALLENGES AND OPPORTUNITIES

A noteworthy challenge for CNN based solutions is the scarcity of publicly available data for training CNN's. The unavailability of data can potentially hinder the progress of deep learning-based solutions in this important field of research. Another important aspect related to the above discussion is that limited data exists that can quantify leather quality based on the observed defects. This potentially important information for the design of such systems is also confined to domain experts and is available in the closed boundaries of the very industries that are in the leather business. Last but not least, an inherent problem at the technical end is to handle leather samples that exhibit a high degree of variance in terms of the defects they contain. We can summarize the above points to list the following challenges that may hinder the progress of deep learning methods in this scintillating field of research.

- how to collect ample data and capture all the defect variations for robust training of convolutional neural networks?
- how to define a quantitative measure for leather quality based on observed defects?
- how to deal with data samples having a high degree of variance in terms of defects (detecting several defects simultaneously)?

Fortuitously, we can turn these challenges into opportunities for future research. Specifically, we can leverage effective data augmentation techniques for ample data generation and develop automated annotation methods. Devise automated methods for quantification of leather quality and research design/development of CNN based solutions for simultaneous multi-defect inspection. In the coming section, we provide potential future research directions for handling these challenges.

B. FUTURE RESEARCH DIRECTIONS

1) DATA AUGMENTATION

For automatic data generation, an important direction of future research will be to devise unique data augmentation methods for data generation. Also, it will be vital to develop automated annotation methods for facilitating weakly supervised learning as manual annotation would not be feasible on such a big scale.

2) QUANTIFYING LEATHER QUALITY

Provided ample grading data from domain experts, there is a need to design methods that can interpolate on that data and learn to quantify leather quality. This problem can be solved with effective data augmentation techniques, and especially generative adversarial networks may prove to be quite useful for this problem, due to their recent success in similar settings [76]. According to this proposal, the generator network generates a rating for the leather quality, where the discriminator network attempts to discriminate it in terms of whether it is from the human expert or the generator network.

3) SIMULTANEOUS DEFECT INSPECTION

To counter the challenge of simultaneous multi-defect detection, it is vital to resort to CNNs, owing to their inherent capabilities and success in recent literature. CNN's are an inherent choice for simultaneous detection of multi-class categories, which is considered a fact after their triumph on large scale data having thousands of classes. Hence, the design and development of CNN based machine vision systems for robust visual inspection of leather and hides is an important future research direction.

For challenging cases, where a leather sample contains several different types of a defect having a high degree of variability, the traditional CNN based methods may not obtain the best results. For industrially applicable deep learning in such scenarios, the problem of defect detection may be decoupled from the problem of leather quality grading. This can be achieved through an ensemble of CNNs.

IX. CONCLUSION

This paper presented a methodical and detailed review on machine vision for leather defect inspection and grading. A detailed review of the image analysis based leather defect inspection methods that generally belong to the class of heuristic or basic machine learning techniques was presented. Owing to the recent success of deep learning methods in various related intelligent applications, deep learning architectures tailored to image classification, detection and segmentation are discussed. A detailed review of the role of deep learning methods in general visual inspection applications was presented, where recent CNN architectures are classified and compared. The recent triumph of CNN based methods for general defect inspection is a driver for their application in leather defect detection/classification. Also, these CNN architectures can act as a source of guidelines for the design and development of novel solutions for leather defect inspection.

In this work, we highlighted the challenges that exist in the design and development of CNN based solutions for leather defect inspection, where ample training data, quantification of leather quality and high variability of defects are some of the greatest challenges. We also presented research directions for fellow researchers that should be investigated in the future to overcome these challenges and enable advancements in this very important area of research.

REFERENCES

[1] D. Mukesh and K. A. Kumar, "Industrial examples," in *Green Chemistry and Engineering*. Burlington, VT, USA: Academic, 2007, ch. 9, pp. 245–296.

[2] P. Masciana, *World Statistical Compendium for Raw Hides and Skins, Leather and Leather Footwear*. Rome, Italy: Food and Agriculture Organization, 2015, pp. 1–122.

[3] B. Antonella, T. Maria, A. Giovanni, and D. Arcangelo, "Automated system for detection and classification of leather defects," *Opt. Eng.*, vol. 35, no. 12, pp. 3485–3494 and 10, 1996.

[4] A. F. Limas-Serafim, "Natural images segmentation for patterns recognition using edges pyramids and its application to the leather defects," in *Proc. 9th Annu. Conf. IEEE Ind. Electron. (IECON)*, Nov. 1993, pp. 1357–1360.

[5] F. P. Lovergine, A. Branca, G. Attolico, and A. Distante, "Leather inspection by oriented texture analysis with a morphological approach," in *Proc. Int. Conf. Image Process.*, vol. 2, Oct. 1997, pp. 669–671.

[6] A. Branca, F. P. Lovergine, G. Attolico, and A. Distante, "Defect detection on leather by oriented singularities," in *Proc. Int. Conf. Comput. Anal. Images Patterns*. Berlin, Germany: Springer, 1997, pp. 223–230.

[7] A. Kumar and G. K. H. Pang, "Defect detection in textured materials using Gabor filters," *IEEE Trans. Ind. Appl.*, vol. 38, no. 2, pp. 425–440, Mar. 2002.

[8] J. L. Sobral, "Leather inspection based on wavelets," in *Pattern Recognition and Image Analysis*, J. S. Marques, N. P. de la Blanca, and P. Pina, Eds. Berlin, Germany: Springer, 2005, pp. 682–688.

[9] F. Q. He, W. Wang, and Z. C. Chen, *Automatic Visual Inspection for Leather Manufacture*, vol. 326. Zürich, Switzerland: Trans Tech Publication, 2006, pp. 469–472.

[10] M. Jawahar, N. K. C. Babu, and K. Vani, "Leather texture classification using wavelet feature extraction technique," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res.*, Dec. 2014, pp. 1–4.

[11] R. M. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE*, vol. 67, no. 5, pp. 786–804, May 1979.

[12] H. Pistori, W. A. Paraguassu, P. S. Martins, M. P. Conti, M. A. Pereira, and M. A. Jacinto, "Defect detection in raw hide and wet blue leather," in *Proc. Int. Symp. CompIMAGE*, 2018, p. 355.

[13] R. Viana, R. B. Rodrigues, M. A. Alvarez, and H. Pistori, "SVM with stochastic parameter selection for bovine leather defect classification," in *Advances in Image and Video Technology*, D. Mery and L. Rueda, Eds. Berlin, Germany: Springer, 2007, pp. 600–612.

[14] C. Yeh and D.-B. Perng, "A reference standard of defect compensation for leather transactions," *Int. J. Adv. Manuf. Technol.*, vol. 25, nos. 11–12, pp. 1197–1204, 2005.

[15] L. Jian, H. Wei, and H. Bin, "Research on inspection and classification of leather surface defects based on neural network and decision tree," in *Proc. Int. Conf. Comput. Design Appl.*, 2010, p. V2-381.

[16] P. Villar, M. Mora, and P. Gonzalez, *A New Approach for Wet Blue Leather Defect Segmentation*. Berlin, Germany: Springer, 2011, pp. 591–598.

[17] R. F. Pereira, C. M. S. Medeiros, and P. P. R. Filho, "Goat leather quality classification using computer vision and machine learning," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2018, pp. 1–8.

[18] W. P. Amorim, H. Pistori, M. C. Pereira, and M. A. C. Jacinto, "Attributes reduction applied to leather defects classification," in *Proc. 23rd SIBGRAPI Conf. Graph., Patterns Images*, 2010, pp. 353–359.

[19] A. Lerch and D. Chetverikov, "Knowledge-based line-correction rules in a machine-vision system for the leather industry," *Eng. Appl. Artif. Intell.*, vol. 4, no. 6, pp. 433–438, 1991.

[20] W. Poelzleitner and A. Niel, *Automatic Inspection of Leather Surfaces*. vol. 2347. Boston, MA, USA: International Society for Optics and Photonics, 1994, pp. 50–59.

[21] K. Krastev, L. Georgieva, and N. Angelov, "Leather features selection for defects' recognition using fuzzy logic," *Energy*, vol. 2, p. 3, Jun. 2004.

[22] C. Yeh and D.-B. Perng, "Establishing a demerit count reference standard for the classification and grading of leather hides," *Int. J. Adv. Manuf. Technol.*, vol. 18, no. 10, pp. 731–738, 2001.

[23] C. Kwak, J. A. Ventura, and K. Tofang-Sazi, "Automated defect inspection and classification of leather fabric," *Intell. Data Anal.*, vol. 5, no. 4, pp. 355–370, 2001.

[24] L. Georgieva, K. Krastev, and N. Angelov, "Identification of surface leather defects," in *Proc. 4th Int. Conf. Conf. Comput. Syst. Technol., e-Learn.*, 2003, pp. 303–307.

[25] M. K. Kasi, J. B. Rao, and V. K. Sahu, "Identification of leather defects using an autoadaptive edge detection image processing algorithm," in *Proc. Int. Conf. High Perform. Comput. Appl. (ICHPCA)*, 2014, pp. 1–4.

[26] C.-C. Ho, J.-C. Li, T. H. Kuo, and C.-C. Peng, "Multicamera fusion-based leather defects marking system," *Adv. Mech. Eng.*, vol. 5, pp. 1–7, Jan. 2013.

[27] J. D. A. Penaranda, J. A. R. Alcazar, L. M. T. Balibrea, J. L. M. Lozano, and R. T. Sanchez, "Inspection and measurement of leather system based on artificial vision techniques applied to the automation and waterjet cut direct application," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, vol. 1, Oct. 1994, pp. 863–867.

[28] S. Anand, C. McCord, R. Sharma, and T. Balachander, "An integrated machine vision based system for solving the nonconvex cutting stock problem using genetic algorithms," *J. Manuf. Syst.*, vol. 18, no. 6, pp. 396–415, 1999.

[29] J. Paakkari, H. J. Ailisto, M. Niskala, M. Makarainen, and K. Vainamo, *Machine-Vision-Guided Waterjet Cutting*, vol. 3827. Munich, Germany: International Society for Optics and Photonics, 1999, pp. 44–52.

[30] M. Lanzetta and G. Tantussi, *Design and Development of a Vision Based Leather Trimming Machine*. Vienna, Austria: Springer, 2002, pp. 561–568.

[31] A. F. L. Serafim, "Multiresolution pyramids for segmentation of natural images based on autoregressive models: Application to calf leather classification," in *Proc. Int. Conf. Ind. Electron., Control Instrum.*, 1991, pp. 1842–1847.

[32] M. Jawahar, K. Vani, and N. Babu, "Machine vision inspection system for detection of leather surface defects," *J. Amer. Leather Chemists Assoc.*, vol. 114, pp. 10–19, Jan. 2019.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:0706.1234*. [Online]. Available: https://arxiv.org/abs/arXiv:0706.1234

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/arXiv:1409.1556

[35] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[36] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[37] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," 2017, *arXiv:1702.05659*. [Online]. Available: https://arxiv.org/abs/arXiv:1702.05659

[38] Y. Tang, "Deep learning using linear support vector machines," 2013, *arXiv:1306.0239*. [Online]. Available: https://arxiv.org/abs/arXiv:1306.0239

[39] W. M. Czarnecki, R. Jozefowicz, and J. Tabor, "Maximum entropy linear manifold for learning discriminative low-dimensional representation," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Cham, Switzerland: Springer, 2015, pp. 52–67.

[40] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[41] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2015, pp. 91–99.

[43] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Apr. 2013.

[44] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–10.

[45] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 21–37.

[46] Z. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[47] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.

[48] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[49] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[50] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham, Switzerland: Springer, 2015, pp. 234–241.

[51] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[52] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *CoRR*, vol. abs/1706.05587, p. 7, Jun. 2017.

[53] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. 15th Eur. Conf. Comput. Vis. ECCV*, Munich, Germany, Sep. 2018, pp. 833–851.

[54] L.-C. Chen, J. T. Barron, G. Papandreou, K. Murphy, and A. L. Yuille, "Semantic image segmentation with task-specific edge detection using cnns and a discriminatively trained domain transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4545–4554.

[55] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[56] S. Dorafshan, R. J. Thomas, and M. Maguire, "Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete," *Construct. Building Mater.*, vol. 186, pp. 1031–1045, Oct. 2018.

[57] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, and A. Agrawal, "Deep convolutional neural networks with transfer learning for computer vision-based data-driven pavement distress detection," *Construction Building Mater.*, vol. 157, pp. 322–330, Dec. 2017.

[58] F. Chen and M. R. Jahanshahi, "NB-CNN: Deep learning-based crack detection using convolutional neural network and Naïve Bayes data fusion," *IEEE Trans. Ind. Electron.*, vol. 65, no. 5, pp. 4392–4400, May 2018.

[59] Y. Li, W. Zhao, and J. Pan, "Deformable patterned fabric defect detection with Fisher criterion-based deep learning," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 2, pp. 1256–1264, Apr. 2017.

[60] S. Mei, Y. Wang, and G. Wen, "Automatic fabric defect detection with a multi-scale convolutional denoising autoencoder network model," *Sensors*, vol. 18, no. 4, p. 1064 Apr. 2018.

[61] J. Chen, Z. Liu, H. Wang, A. Nunez, and Z. Han, "Automatic defect detection of fasteners on the catenary support device using deep convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 2, pp. 257–269, Feb. 2018.

[62] M. Ferguson, R. Ak, Y.-T. Lee, and K. H. Law, "Automatic localization of casting defects with convolutional neural networks," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 1726–1735.

[63] D. Weimer, B. Scholz-Reiter, and M. Shpitalni, "Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection," *CIRP Ann.-Manuf. Technol.*, vol. 65, no. 1, pp. 417–420, 2016.

[64] Z. Yu, X. Wu, and X. Gu, "Fully convolutional networks for surface defect inspection in industrial environment," in *Computer Vision Systems*, M. Liu, H. Chen, and M. Vincze, Eds. Cham, Switzerland: Springer, 2017, pp. 417–426.

[65] S. Faghih-Roohi, S. Hajizadeh, A. Núñez, R. Babuska, and B. De Schutter, "Deep convolutional neural networks for detection of rail surface defects," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2016, pp. 2584–2589.

[66] X. Gibert, V. M. Patel, and R. Chellappa, "Deep multitask learning for railway track inspection," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 153–164, Jan. 2017.

[67] G. Psuj, "Multi-sensor data integration using deep learning for characterization of defects in steel elements," *Sensors*, vol. 18, no. 1, p. 292, 2018.

[68] J. Masci, U. Meier, D. Ciresan, J. Schmidhuber, and G. Fricout, "Steel defect classification with max-pooling convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2012, pp. 1–6.

[69] R. Ren, T. Hung, and K. C. Tan, "A generic deep-learning-based approach for automated surface inspection," *IEEE Trans. Cybern.*, vol. 48, no. 3, pp. 929–940, Mar. 2018.

[70] J. Feng, F. Li, S. Lu, J. Liu, and D. Ma, "Injurious or noninjurious defect identification from MFL images in pipeline inspection using convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 66, no. 7, pp. 1883–1892, Jul. 2017.

[71] Y.-J. Cha, W. Choi, G. Suh, S. Mahmoudkhani, and O. Büyüköztürk, "Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types," *Comput.-Aided Civil Infrastruct. Eng.*, vol. 33, no. 9, pp. 731–747, 2018.

[72] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[73] *Gpec—Research Group in Engineering and Computing*. Accessed: Jun. 11, 2019. [Online]. Available: http://www.gpec.ucdb.br/ingles/index.html

[74] S.-T. Liong, Y. S. Gan, Y.-C. Huang, C.-A. Yuan, and H.-C. Chang, "Automatic defect segmentation on leather with deep learning," 2019, *arXiv:1903.12139*. [Online]. Available: https://arxiv.org/abs/arXiv:1903.12139

[75] H.-Q. Bong, Q.-B. Truong, H.-C. Nguyen, and M.-T. Nguyen, "Vision-based inspection system for leather surface defect detection and classification," in *Proc. 5th NAFOSTED Conf. Inf. Comput. Sci. (NICS)*, 2018, pp. 300–304.

[76] J. Son, S. J. Park, and K.-H. Jung, "Retinal vessel segmentation in fundoscopic images with generative adversarial networks," 2017, *arXiv:1706.09318*. [Online]. Available: https://arxiv.org/abs/arXiv:1706.09318

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 FAST-NUCES, Islamabad, Pakistan, in 2018. He is currently working as a Research Associate with the Visual Computing Technology (VC-Tech) Laboratory, Islamabad. Before joining VC-Tech, he worked as a Research Assistant with FAST-NUCES. His research interest includes computer vision and image processing.

TARIQ M. KHAN received the Ph.D. degree from Macquarie University, Sydney, Australia. He has over ten years of research and teaching experience in universities, including CUI, Macquarie University, and UET Taxila, Pakistan. He is currently an Assistant Professor with COMSATS University Islamabad (CUI), Pakistan. At CUI, he is also serving as the Head of the Image and Video Processing Research Group (IVPRG). He is interested in both digital image processing (with an emphasis on medical imaging) and machine learning. He has published 34 journals and 17 conference papers in well-reputed journals and conferences, e.g., the IEEE TIP, *Pattern Recognition*, IEEE Access, *Expert Systems with Applications*, JRTIP, ICIAR, VCIP, and DICTA. His research interests include most aspects of image enhancement, pattern recognition, and image analysis. His research has received research funding from Effat University, HEC, Pakistan, MBIE, New Zealand, and KSA International Collaboration Grant over 1M in past years. He is an active member of IEEE for over seven years. He was a recipient of several awards, including iMQRES scholarship, Macquarie University Postgraduate Research Student Support (PRSS) travel grants, and SRGP.

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. His research interests include saliency modelling, 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 HND in leather technology from the Institute for Creative Leather Technologies, University of Northampton, in 1989. He has over 30 years' experience in the leather industry in various production and technical roles, and for the last 14 years, he has been involved in research at the New Zealand Leather and Shoe Research Association (LASRA). He is currently the Director of LASRA. He is also a leather technologist.

RAFEA NAFFA received the Ph.D. degree in biochemistry from Massey University, in 2017. He is currently a Scientist with the New Zealand Leather and Shoe Research Association (LASRA), New Zealand. His research interest focuses on studying collagen molecular composition and its structure in skin and leather using analytical and structural techniques.