ISSN 2349-4506
Impact Factor: 3.799

Global Journal of Engineering Science and Research Management

WHEN IN-DOMAIN IMAGE RECOGNITION FAILS : A CONTROLLED DOMAIN-SHIFT STUDY OF HANDCRAFTED FEATURES FOR STEEL-SURFACE DEFECTS

Dr. Vijay R. Rathod
Head, Department of Electronics and Telecommunication Engineering
ST XAVIERS TECHNICAL INSTITUTE MAHIM MUMBAI 400016
Email id: vijay.r@xaviertech.ac.in

KEYWORDS: Industrial Inspection; Surface-Defect Recognition; Domain Shift; Local Binary Pattern; HOG; Synthetic Benchmark

ABSTRACT

Industrial image-recognition models are often evaluated on randomly partitioned images generated under one texture and illumination process. This study quantifies how that protocol can misrepresent robustness. A reproducible 64×64 grayscale benchmark was procedurally generated with five balanced classes: clean surface, scratch, pits, inclusion, and scale. Training comprised 900 images; independent in-domain and shifted-domain tests contained 400 images each. The shifted set changed background orientation, noise level, and defect contrast. Four standardized linear-SVM pipelines were compared: downsampled intensities, histogram of oriented gradients (HOG), rotation-invariant local binary patterns (RI-LBP), and combined HOG+LBP. In-domain, the combined representation achieved the best macro-F1 of 0.577. Under texture shift, it collapsed to 0.101, while the simpler intensity baseline produced the highest shifted macro-F1 of only 0.301. RI-LBP retained 28.8% accuracy but macro-F1 was 0.238, indicating uneven class transfer. A severity stress test showed defect macro-F1 increasing from 0.245 for low-contrast anomalies to 0.358 for high-contrast anomalies. The negative result demonstrates that feature fusion can amplify source-texture dependence and that synthetic validation must deliberately change the generating process. These values are diagnostic of the controlled benchmark, not estimates for factory inspection; real NEU or MVTec validation is required.

KEYWORDS - Industrial Inspection; Surface-Defect Recognition; Domain Shift; Local Binary Pattern; HOG; Synthetic Benchmark

INTRODUCTION

Machine-vision inspection promises consistent, high-speed identification of scratches, pits, inclusions, scale, and other surface anomalies. Public resources such as the NEU steel defect database and MVTec AD have made algorithms easier to compare [1–3]. Yet a random train–test split can retain nearly identical texture, illumination, and acquisition signatures in both partitions. A classifier may therefore recognize the image generator or production batch rather than a transferable defect concept.

The risk is particularly relevant for handcrafted descriptors. HOG encodes oriented edges; LBP encodes local intensity order. Both are efficient, but a change in background direction, sensor noise, or anomaly contrast can alter their distributions [4,5]. Combining descriptors may increase in-domain separability while also increasing sensitivity to source conditions. An evaluation that never changes the data-generating process cannot expose this trade-off.

This study creates a small controlled benchmark inspired by the synthetic-domain logic of the DAGM 2007 inspection competition [6]. It asks: (i) which representation performs best when training and testing share a generator; (ii) how strongly performance changes when the texture generator is shifted; and (iii) whether low-contrast anomalies are disproportionately difficult. The purpose is methodological stress testing, not a substitute for a real steel dataset.

CONTEXT AND BENCHMARK DESIGN

2.1 Relation to public industrial datasets

The NEU resource contains 1,800 grayscale steel images across six defect classes and explicitly notes large within-class variation, between-class similarity, illumination changes, and material changes [1]. Song and Yan showed the relevance of completed LBP variants for noise-robust recognition [2]. MVTec AD broadened industrial anomaly benchmarking to more than 5,000 images from 15 object and texture categories with pixel-level anomaly masks [3]. DAGM's competition used artificially generated textures and tested on texture and defect models that

were unknown during development [6]. The present benchmark follows that controlled philosophy at a smaller classification scale.

2.2 Procedural image generator

Each image contained a brushed grayscale background formed from an oriented sinusoidal component, smoothed low-frequency random variation, and pixel noise. Five labels were generated: clean; scratch (one or two thin oriented lines); pits (multiple dark Gaussian depressions); inclusion (one irregular compact region); and scale (thresholded low-frequency patches). Values were clipped to [0,1]. The generator produced 180 training images per class (900 total), plus 80 images per class for each test domain (400 in-domain and 400 shifted).

Training and in-domain backgrounds used orientations from −12° to 18°. The shifted domain used 25° to 55°, increased pixel-noise standard deviation from 0.035 to 0.050, and reduced nominal defect contrast to 80% of the training level. Independent seeds generated every partition. These changes were selected before fitting and apply to all classes; they do not use test labels to modify a classifier.

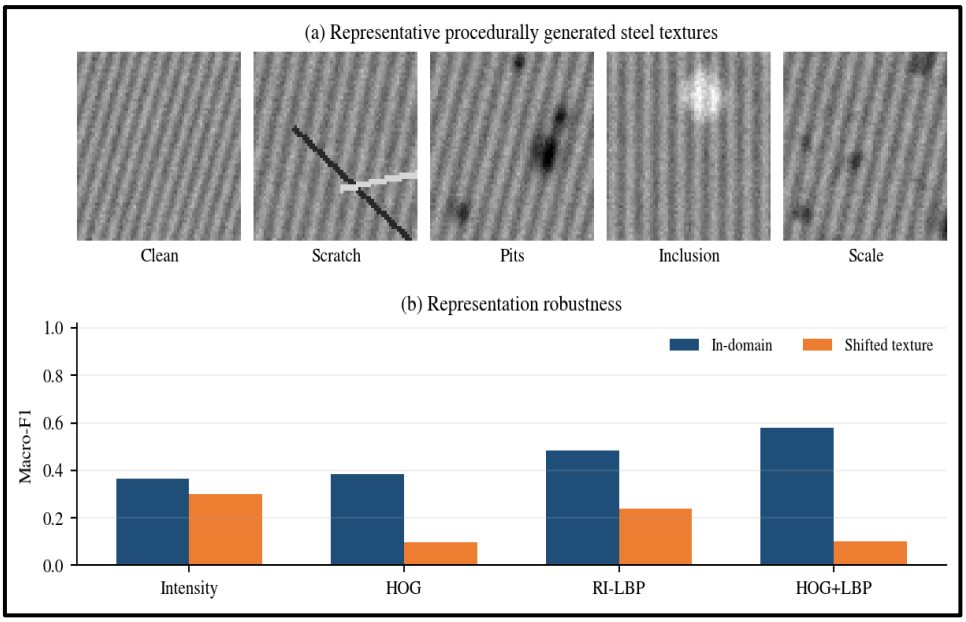

Figure 1. Representative procedurally generated surfaces and representation performance across domains.

FEATURE PIPELINES AND EVALUATION

3.1 Image representations

The intensity baseline resized each 64×64 image to 16×16 and flattened 256 normalized values. HOG used unsigned gradients, nine orientation bins, 8×8-pixel cells, 2×2-cell overlapping blocks, and L2-Hys normalization [5]. RI-LBP compared each centre pixel with its eight neighbours, selected the minimum circular bit rotation, and accumulated normalized code histograms over a 4×4 spatial grid [4]. The combined representation concatenated HOG and RI-LBP.

3.2 Classifier and metrics

Every representation was standardized using training-set means and standard deviations and classified with the same one-versus-rest linear SVM (C=0.7; maximum 10,000 iterations). Holding the classifier fixed isolates representation behaviour. Accuracy, balanced accuracy, and macro-F1 were computed for both domains. Because all classes were balanced, accuracy and balanced accuracy were numerically identical. Image-level 95% confidence intervals for accuracy used 500 bootstrap resamples.

ISSN 2349-4506
Impact Factor: 3.799
Global Journal of Engineering Science and Research Management

Pipeline	Dimensional idea	Expected strength	Expected vulnerability
Intensity	16×16 spatial values	Coarse morphology	Position and photometric shift
HOG	Local gradient blocks	Edges and scratches	Background orientation
RI-LBP	Rotation-normalized code histograms	Micro-texture	Noise near intensity ties
HOG+LBP	Concatenated edge and texture cues	In-domain complementarity	Combined source dependence

3.3 Defect-severity stress test

The pipeline with the highest shifted-domain macro-F1 was locked and evaluated on three additional shifted sets containing 450 images each. Defect contrast multipliers were 0.40 (low), 0.65 (medium), and 0.95 (high). The clean class was retained for overall accuracy, while defect macro-F1 was computed only across scratch, pits, inclusion, and scale. This analysis tests visibility, not calibration.

Chance accuracy for five balanced classes is 0.20. Macro-F1 can fall below 0.20 when predictions collapse into a subset of labels, making it more informative than accuracy for diagnosing asymmetric transfer.

RESULTS

4.1 In-domain versus shifted-domain recognition

Table 1. Performance by representation and test domain; accuracy and interval are percentages.

Representation	Domain	Accuracy	95% interval	Macro-F1
Downsampled intensity	In-domain	37.0	32.4–41.8	0.367
Downsampled intensity	Shifted	28.2	24.0–33.0	0.301
HOG	In-domain	38.8	34.0–42.9	0.384
HOG	Shifted	20.5	16.8–24.4	0.096
Rotation-invariant LBP	In-domain	48.0	43.2–52.4	0.485
Rotation-invariant LBP	Shifted	28.7	24.2–33.2	0.238
HOG + LBP	In-domain	57.2	52.4–62.4	0.577
HOG + LBP	Shifted	21.5	17.8–25.8	0.101

In-domain, concatenating HOG and RI-LBP produced the strongest result (57.3% accuracy; macro-F1 0.577), followed by RI-LBP (48.0%; 0.485). The combined representation did not transfer: shifted accuracy fell to 21.5% and macro-F1 to 0.101. HOG alone showed the same collapse (macro-F1 0.096). The downsampled-intensity baseline was weak in-domain (0.367 macro-F1) but highest under shift (0.301), although its 28.3% accuracy remained only modestly above chance. RI-LBP achieved 28.8% shifted accuracy but a lower macro-F1 of 0.238, indicating that its correct predictions were not evenly distributed across classes.

4.2 Visibility and confusion

Table 2. Shifted-domain severity stress test for the locked intensity pipeline.

Defect contrast	Overall accuracy %	Defect-only macro-F1
Low	22.4	0.245
Medium	24.2	0.271
High	29.6	0.358

Defect-only macro-F1 increased monotonically from 0.245 at low contrast to 0.358 at high contrast. Thus, the best shifted pipeline still relied strongly on anomaly visibility. Its class-level confusion matrix shows that errors were distributed across visually diffuse classes rather than being explained by one isolated label.

ISSN 2349-4506
Impact Factor: 3.799
Global Journal of Engineering Science and Research Management

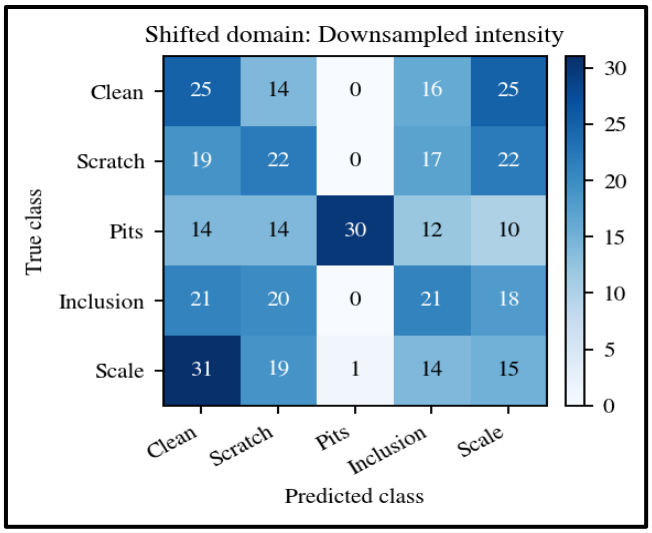

Figure 2. Shifted-domain confusion matrix for the representation with the highest shifted macro-F1.

DISCUSSION

The experiment demonstrates that the representation ranked first under an in-domain test can be among the worst after a controlled generator shift. HOG+LBP benefited from complementary edge and texture cues when the background orientation matched training. Under shift, the high-dimensional concatenation carried source-specific structure into the linear decision function. HOG was especially vulnerable because its orientation histograms treated the altered brushing direction as a dominant change. RI-LBP was designed to reduce rotational sensitivity, yet its local comparisons remained sensitive to increased noise and lower defect contrast [2,4].

The intensity baseline's relative advantage under shift should not be celebrated as strong recognition. Its shifted macro-F1 of 0.301 is inadequate for inspection and only slightly above chance. Rather, it indicates that aggressive downsampling discarded some source texture while preserving coarse defect morphology. This is a useful diagnostic result: under domain shift, a simpler representation may fail less severely even when it is inferior in-domain.

5.1 Implications for experimental design

Random image splits from a single production batch are insufficient evidence of deployment readiness. Industrial studies should partition by coil, production day, camera, illumination configuration, or material grade whenever those identifiers exist. If only synthetic data are available, the evaluation generator should differ from the training generator, as in the DAGM competition design [6]. Model selection must occur without access to the shifted test labels, and class-specific errors should be reported alongside aggregate accuracy.

5.2 Limitations

The generator is intentionally simple and does not reproduce hot-rolled steel physics, specular reflection, scale-dependent optics, motion blur, nonuniform lighting, or annotation ambiguity. Its five classes are inspired by common defect morphologies but are not direct copies of NEU images. Consequently, the numerical performance cannot be compared with NEU or MVTec leaderboards. The stress test changes several factors together, so it does not identify the separate causal contribution of orientation, noise, or contrast. Linear SVMs and handcrafted descriptors were chosen for interpretability and speed; convolutional, self-supervised, and anomaly-detection models were outside scope.

Future work should preregister a factorial shift design, validate on NEU-CLS and MVTec texture categories [1,3], use group-wise splits, and quantify calibration and false-negative cost. Synthetic pretraining should be assessed against real-only and mixed-data baselines at equal labelling budgets. Localization masks would also test whether a classifier attends to the defect rather than the background.

CONCLUSIONS

A controlled change in texture orientation, noise, and defect contrast reversed the apparent ranking of handcrafted steel-defect representations. HOG+LBP reached the best in-domain macro-F1 (0.577) but collapsed to 0.101 under

shift; the simple intensity baseline retained the highest shifted macro-F1, only 0.301. Lower defect contrast further reduced performance. The negative result is the contribution: in-domain image recognition is not evidence of process-level robustness, and synthetic benchmarks should deliberately alter their generating mechanisms.

REFERENCES

1. Song K, Yan Y. NEU surface defect database. Northeastern University. https://faculty.neu.edu.cn/songkc/en/zdylm/263265/.
2. Song K, Yan Y. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. Applied Surface Science. 2013;285:858–864. doi:10.1016/j.apsusc.2013.09.002.
3. Bergmann P, Fauser M, Sattlegger D, Steger C. MVTec AD—A comprehensive real-world dataset for unsupervised anomaly detection. IEEE CVPR. 2019:9584–9592. doi:10.1109/CVPR.2019.00982.
4. Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE TPAMI. 2002;24(7):971–987. doi:10.1109/TPAMI.2002.1017623.
5. Dalal N, Triggs B. Histograms of oriented gradients for human detection. IEEE CVPR. 2005:886–893. doi:10.1109/CVPR.2005.177.
6. Wieler M, Hahn T. Weakly supervised learning for industrial optical inspection. DAGM 2007 Competition. https://conferences.mpi-inf.mpg.de/dagm/2007/prizes.html.
7. He Y, Song K, Meng Q, Yan Y. An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. IEEE Transactions on Instrumentation and Measurement. 2020;69(4):1493–1504. doi:10.1109/TIM.2019.2915404.
8. Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;20:273–297. doi:10.1007/BF00994018.
9. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.