https://doi.org/10.2352/EI.2024.36.9.IQSP-263

A Novel Deep Learning-Based Anomaly Detection Pipeline Optimized for Dark Blemish Pattern of Image Sensor

Dayun Lee, Yubin Lee, Dagyeom Hong, Kundong Kim, Sung-su Kim, and Yitae Kim; S.LSI Division, Samsung Electronics; Hwaseong-si, Gyeonggi-do, Republic of Korea

Abstract

Nowadays, the quality of low-light pictures is becoming a competitive edge in mobile phones. To ensure this, the necessity to filter out dark defects that cause abnormalities in dark photos in advance is emerging, especially for dark blemish. However, high manpower is required to separate dark blemish patterns due to the low consistency problem of the existing scoring method. This paper proposes a novel deep learning-based screening method to solve this problem. The proposed pipeline uses two ResNet-D models with different depths to perform classification and regression of visibility, respectively. Then it derives a new score that combines the outputs of both models into one. In addition, we collect the large-scale image set from real manufacturing processes to train models and configure the dataset with two types of label systems suitable for each model. Experimental results show the performance of the deep learning models trained and validated with the presented datasets. Our classification model has significantly improved screening performance with respect to its accuracy and F1-score compared to the conventional handcraft method. Also, the visibility regression method shows a high Pearson correlation coefficient with 30 expert engineers, and the inference output of our regression model is consistent with it.

Introduction

With the development of smartphone camera performance, the quality of pictures captured in low-light conditions is becoming a competitive edge. Consequently, more and more phone settings are adopting high gain for low-light shooting. However, the high gain is applied not only to pure signals but also to defects and noises. As a result, very small defective signal may be amplified and recognizable. We call these kinds of defects, which are problematic in dark conditions, as dark defects. *Dark blemish* is one kind of these dark defects. It is generated during the manufacturing process and has a random shape and position. If we use an image sensor module with dark blemish, your picture of the night sky may show some stain, which significantly drops the image quality. Therefore, it is important to screen modules containing dark blemishes in advance. Figure 1 shows examples of amplified dark images and a picture of the night sky from a module with dark blemish patterns.

There is an existing method to screen dark blemishes using image processing algorithms. First, the dark image is enhanced so that the hidden patterns can be seen. Then the patterns are separated and scored by subtracting the min value of the pattern from the max value of the pattern. After that, the sets of scores and enhanced images are passed to engineers. Engineers inspect them and decide the threshold at which the dark blemish could be visible on the display of phones. We call this threshold *spec*. However, this conventional method has a few problems. First, the score has a low correlation to actual visibility. Because the enhanced dark images are noisy, so it is hard to separate and score pure patterns only. Second, the range of score distribution varies by product because of the different features of them. (Figure 2) This leads to the third problem that engineers have to repeat all inspection process and decide the specs for every product. This handcrafted method is inefficient since it requires a lot of manpower and results may vary depending on the experience of the engineers. [1]

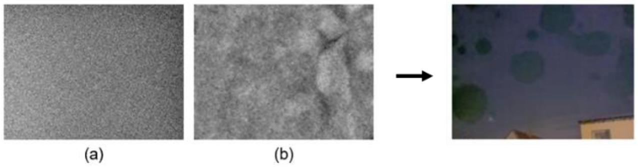

Figure 1. *Examples of amplified dark images. (a) Normal image which only has temporal noise. (b) Defective image with dark blemish. Low-light picture taken by the sensor with dark blemish may show stains.*

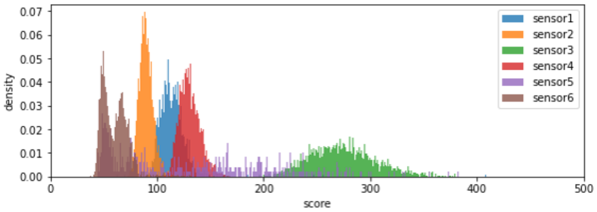

Figure 2. *Distribution of scores by sensors. Most of them have Gaussian-like distributions but with different means and variances*

To handle these problems, we present a deep learning pipeline for dark blemish detection in this paper. Figure 3 represents the suggested pipeline. It is composed of two deep learning models, a classification model and a visibility regression model, and one decision rule. The classification model detects abnormal patterns in images based on shape information, and the visibility regression model scores the pattern's intensity. Then the decision rule is applied to combine two outputs from each model into a final decision.

In method section, we will explain how we compose a large-scale dataset to train and evaluate presented pipeline and how the models are trained. Then in the result section, we will show the performance of suggested models compared to conventional method.

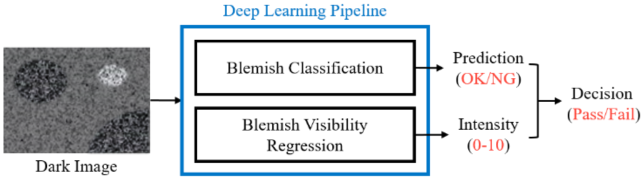

Figure 3. *Proposed deep learning pipeline for dark blemish detection. There are two deep learning models and a final decision.*

Method

Dataset Construction

To train and evaluate the proposed deep learning pipeline, we construct datasets with images taken in the electrical die sorting (EDS) stage where tests are conducted during the manufacturing process. It is known that data imbalance is a common problem in defect datasets due to the properties of mass production. However, we need well-balanced and clean class-image dataset to achieve high performance with a deep learning model. [2] Therefore, we gathered high-score images especially to get enough defective data. The obtained raw images were pre-processed to enhance the visibility of defects by stretching. The enhanced images are used to make both the classification dataset and the regression dataset.

Classification Dataset

We perform a 2-stage classification for our dataset. First, the images are divided into OK and NG. OK corresponds to normal images, and NG contains images with dark blemish patterns. NG images are subdivided into *Spot, Diagonal,* and *Error* again depending on the shape of the patterns. (Figure 4) The reason we subcategorize defective images is we found that various morphological characteristics in the NG class cannot be sufficiently learned using one class. At the same time, we tried to divide them roughly because too diverse classes may lead to too sparse data per class, which intensifies the data imbalance problem. We experimentally confirmed that using a 4-class scheme for the training phase and summarizing the output into OK and NG for the testing phase works better than leveraging simple 2-classes.

Based on this strategy, we conduct labeling on 40,276 images collected. To give a clean label for defects with a wide spectrum, 3 engineers with background knowledge worked together and cross-validated the labels. The labeled classification dataset is divided into 7:2:1 ratios for training, validation, and testing. The numbers of images for each class and split are shown in Table 1.

Regression Dataset

Instead of using the class labels as it is for our regression dataset, we relabel images with 11 levels based on their intensity of pattern. By giving numerical levels instead of discontinuous classes, it is possible to estimate real number intensity through the regression model. We set the leveling criteria as follows:

- Normal images: 0-2
- Ambiguous images: 3
- Defective images: 4-9
- Capture-failed images: 10

We give a twice longer range for defective images than normal images as we focus on separating how bad the pattern is in detail. Scoring was also carried out by 3 engineers, but with many more iterations to have consistent levels. Nevertheless, it is still complicated to score images relying on cognitive senses. Therefore we verified the levels with a qualitative evaluation by 30 engineers, and the comparison result is shown in the result section.

To maintain some balance in the number of data for each level, 6,514 images, a subset of the whole dataset, are used for the regression dataset.

Deep Learning Model

After the datasets are ready, we build deep learning models. We use ResNet-D [3] structure for both classification and regression models. It was presented in the early phase but still one of the most powerful models. We take different model depth, loss function, and optimization function based on each model's purpose. For the classification model, we use ResNet101-D which is relatively deep, to learn shape information accurately. Weighted cross entropy is adopted as a loss function so that the feature of minor classes can be well-trained despite data imbalance. [4] In particular, we increased the weight of Diagonal class because the number of data is relatively small and their characteristics are not clear. For the regression model, we use ResNet18-D model. We intend our model to learn from integer levels and output the levels in real numbers. Therefore by using a shallow model, we prevent our model from overfitting and get the output level in a general manner. Also, we trained the model with Huber loss [5], which is a robust loss function used for a wide range of regression tasks. The optimization function for each training is selected by comparing various candidates. SGD and Adam show the best performance.

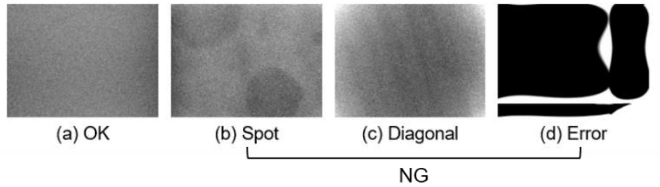

Figure 4. *Sample images of each class*

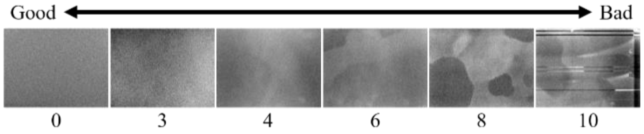

Figure 5. *Sample images of levels*

Table 1. Distribution of Classification Dataset

Split	Total	OK	Spot	Diag.	Error
Train	28,201	24,343	3,456	161	241
Validation	8,029	6,929	986	45	69
Test	4,046	3,501	485	24	36
Total	40,276	34,773	4927	230	346

Table 2. Distribution of Regression Dataset

Split	Total	Lv. 0	Lv. 1	Lv. 2	Lv. 3	Lv. 4	Lv. 5	Lv. 6	Lv. 7	Lv. 8	Lv. 9	Lv. 10
Train	5,224	835	994	569	284	294	335	264	406	760	289	194
Val	644	100	113	75	29	43	40	50	45	78	48	23
Test	646	93	109	81	40	41	48	37	36	102	30	29
Total	6,514	1,028	1,216	725	353	378	423	351	487	940	367	246

Although we build two separate models for classification and visibility regression, we need to consider both how clear the shape is and how strong the stain is for the final decision comprehensively. Therefore, we combined the output of the classification model and the regression model with the following equations:

$$\mathrm{P}_{\mathrm{clf}}(input|NG) = \sum_{i \in NG} softmax(clf(input)[i]) \quad (1)$$

$$I_{reg} = reg(input)/10 \quad (2)$$

$$Score = P_{\mathrm{clf}}(input|NG) * I_{reg} \quad (3)$$

where clf is the classification model and reg is the regression model. In equation (1), the clarity of shape is presented by the probability that an input is classified as NG. It is calculated by adding up the softmax of the classification output of NG subclasses. Then in equation (2), the intensity of an input image is presented by the output of the regression model. We divide the output level by 10 to represent the intensity in [0, 1] range. Finally, the score for the decision is presented as the product of the probability that the input is NG and the intensity of the input in equation (3). Usually, the output of the deep learning model is deterministic after the training is done, and hard to change unless it is retrained. However, we can control the decision with this score by releasing or tightening the accepting threshold for the score.

Result

In this section, we present the performance of the classification model and the regression model trained with the proposed dataset and strategy.

Classification

Concerning classification performance, we compared our classification model to the conventional method for 6 image sensors. For the conventional method, we used the same specs as used in the production which are different for each sensor. On the other hand, we used one model trained with the proposed dataset and strategy for the classification model. The test was conducted based on our classification dataset which intentionally increased the defect rate. Thus, note that the result does not reflect real mass production yield.

The results are shown in Table 3. It shows that the proposed model has equal or better accuracy and F1-score than then conventional method for all products. On average, we improved 9.14 percentage points for accuracy and 14.6 percentage points for F1-score. These results confirm that the deep learning-based classification models can significantly improve the dark blemish classification performance without image sensor-specific detection criteria.

Table 3. Classification performance

Sensor	# Images	Conventional		Proposed	
		Acc.	F1-score	Acc.	F1-score
A	174	**1.0000**	**1.0000**	**1.0000**	**1.0000**
B	253	**1.0000**	**1.0000**	**1.0000**	**1.0000**
C	48	0.8750	0.9286	**0.9583**	**0.9737**
D	704	0.7884	0.7950	**0.9688**	**0.9753**
E	122	0.8197	0.7963	**1.0000**	**1.0000**
F	646	0.9536	0.9038	**0.9923**	**0.9825**
Total	1,947	0.8937	0.8332	**0.9851**	**0.9792**

Regression

For the regression performance, we did 2-phase verification. We checked whether the levels of the regression dataset are well assigned according to their visibility first. Then we confirmed the regression model can represent engineers' evaluation.

Regression Dataset Verification

We verified if the levels of the regression dataset and the engineers' judgments were consistent. Since conventional scores have low consistency with visibility, we conducted a quality evaluation test with 30 engineers who have experience in image quality verification. Then, we compared the mean opinion score (MOS) from the test and our levels. For the evaluation test, we exclude level 10, which is obvious.

Figure 6 is the summary box-whisker plot of our levels versus MOS with whisker=1.5. We can see that every box is concentrated within the ±1 level range. In the case of level 3, it has relatively long whiskers, which means there were various responses compared to other levels. We interpret this result as being due to our assignment of ambiguous images to level 3. Still, its statistics such as quantiles, mean, and median are well aligned with MOS. Also, the Pearson correlation coefficient (PCC) between the level and the median of MOS is 1, which means a perfect positive correlation, and the PCC between the levels and the average of MOS is 0.99, which also means a high correlation. From this result, we confirmed that our dataset is aligned with the engineers' opinions.

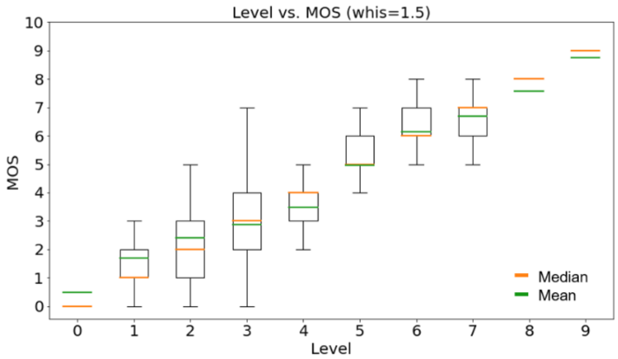

Figure 6. *Level vs. MOS. Each box that represents the range from the first quantile to the third quantile, the median, and the mean of MOS are centered to their level.*

Regression Model Verification

Since we confirmed the authenticity of our regression dataset, we leveraged the dataset to train and test our regression model. The model was trained with the rain split of the regression dataset, and then the mean absolute error (MAE) of the level was measured with the test split of the dataset. Since the levels are designed as integers and our model infers the level as a real number, the MAE within 0.5 is assumed as the correct answer considering rounding.

The test result shows that the MAE of output levels from the test set is 0.67, which is close to 0.5. Also, only 7 images out of 646 test images, which is equivalent to 0.46%, have differences exceeding 3 levels. From these results, we can say that the proposed regression model can decide the level of the image which is aligned with the engineers' evaluation.

Conclusion

For dark blemish detection, we had a few problems with the existing scoring system such as low correlation between conventional score and actual visibility, and varying distribution of scores depending on product, which leads to high human cost to determine the specs. To manage those problems, we propose a deep learning pipeline for screening dark blemish patterns as a solution. We build two models with different purposes, classification and regression, and create datasets with label category suitable for each model. Also, the outputs of the two models are combined into the proposed decision score. Through the experiments, we confirmed that our model can be used as a universal test model with high performance for all products. Moreover, we showed that our model can infer the visibility of defective images which is highly correlated with the level of human cognition through the qualitative test by experts.

References

[1] F. L. Chen, and S. F. Liu, "A neural-network approach to recognize defect spatial pattern in semiconductor fabrication," IEEE Transactions on Semiconductor Manufacturing, vol. 13(3), pp. 366-373, 2000.

[2] J. Kim, et al, "Deep learning based automatic defect classification in through-silicon Via process: FA: Factory automation," 2018 29th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC), IEEE, 2018.

[3] T. He, et al, "Bag of tricks for image classification with convolutional neural networks," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019

[4] M. R. Rezaei-Dastjerdehei, et al, "Addressing imbalance in multi-label classification using weighted cross entropy loss function," in 2020 27th National and 5th International Iranian Conference on Biomedical Engineering (ICBME), IEEE, 2020.

[5] C. Yi, and H. Jian, "Semismooth Newton coordinate descent algorithm for elastic-net penalized Huber loss regression and quantile regression," Journal of Computational and Graphical Statistics, vol. 26(3), pp. 547-557, 2017.

Author Biography

Dayun Lee received her B.S. in electrical engineering from Korea University in 2019, and her M.S. in electrical engineering, computer division from KAIST in 2021. Since then she has worked in Samsung Electronics, Republic of Korea. Her work has focused on the machine learning for image sensors such as automotive tuning and defect classification.