Advances in Manufacturing Technology XXXV
M. Shafik and K. Case (Eds.)

doi:10.3233/ATDE220600

A Survey of Machine Learning Approaches for Visual Inspection on the DAGM Dataset

Philippe CARVALHO[a,1], Alexandre DURUPT[a] and Yves GRANDVALET[b]
[a] *Roberval, Université de technologie de Compiègne, France*
[b] *Heudiasyc, Université de technologie de Compiègne, CNRS, France*

Abstract. The task of automatic visual inspection has been tackled by numerous machine learning algorithms. However, there are no global comparative studies of the performance of these algorithms that use metrics directly relevant to their use in an industrial setting. We survey the performance of machine learning algorithms applied to the DAGM dataset, a reference dataset for industrial visual inspection. Our study reports the performance of 17 algorithms for which the learning and evaluation protocols are clear enough to be reproducible. However, not all of these algorithms are comparable, in the sense that they rely on different labelling or even different data. We group the comparable results, and conclude that the DAGM dataset no longer presents a major difficulty for the algorithms based on knowledge of the location of defects in the images. On the other hand, algorithms using only unlabelled images, which are the easiest to implement in practice, do not yet achieve industrially acceptable performance.

Keywords. Machine learning, Defect detection, Visual inspection

1. Introduction

The best visual inspection systems for automatic defect detection are based on machine learning models such as convolutional neural networks (CNN). It is necessary to evaluate these algorithms in realistic environments because industrial inspection takes place in a very particular context for machine learning algorithms: there are many more non-defective items than defective ones; defects can be very diverse (colour, size, shape), and acquisition conditions (lighting, vibrations, temperature, etc.) can influence image quality. Finally, for an industrial inspection system to be considered usable, it must be accurate, with error rates (false positives and false negatives) in the order of one percent.

Data sets for industrial visual inspection are difficult to collect, and defect labelling must be done by an expert. There are few such datasets available today: the DAGM dataset [1], published in 2007, is the historical reference in the field. This dataset, whose properties are adapted to industrial realities, can therefore be used to compare visual inspection approaches. Here we review the results obtained so far on this dataset.

Section 2 presents the characteristics of the DAGM dataset. Section 3 presents the performance criteria we have chosen here, and the methodology for collecting all the

[1] Corresponding Author. philippe.carvalho@utc.fr

necessary information from the bibliography. Section 4 proposes a discussion of the results and concludes with our final remarks.

2. The DAGM Dataset

The DAGM dataset[2] [1] was published as part of the 2007 DAGM symposium for a competition on the subject of "weakly supervised learning for industrial optical inspection", in order to help improve visual defect detection algorithms in industrial settings. The dataset was generated artificially, but was engineered to resemble real defect detection problems in the manufacturing industry. It is divided in two parts. The first part, containing 6 classes, is commonly referred to as the development dataset. The second part contains 4 classes and is referred to as the competition dataset. Examples of defective samples from each class are presented in Figure 1.

Each development dataset contains 1000 images without defect and 150 images with defects. Each competition dataset contains 2000 images without defect and 300 images with defect. All images are 512x512 pixel 8-bit grayscale (pixel values range from 0 to 255). Each dataset is also divided into a training and a testing sub-dataset. They each contain half of the data, but the number of defective and non-defective items is not balanced between both sets. Note that a deprecated link to the dataset[3] does not include the training and testing sets and thus should be avoided to avoid reproducibility issues.

There is a maximum of one defect per image, each defect is located by an ellipse surrounding its precise location. This dataset has been termed weakly labelled because the ellipses do not trim the defects to the pixel level. This should not be confused with the definition of weak supervision in this article, which will be given in Section 3.

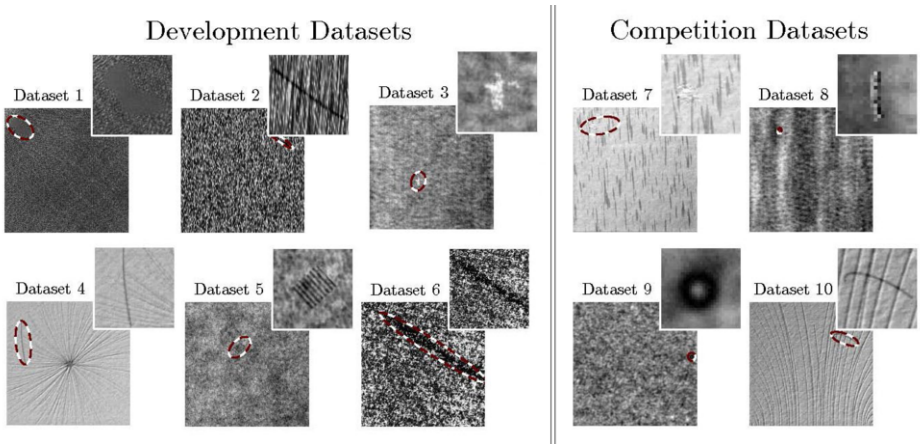

Figure 1. Samples from the DAGM dataset, from [1].

[2] Link to the dataset: https://hci.iwr.uni-heidelberg.de/content/weakly-supervised-learning-industrial-optical-inspection

[3] Deprecated: https://conferences.mpi-inf.mpg.de/dagm/2007/prizes.html

3. Survey of the Performance of Existing Algorithms

The statistics we have chosen to report here are: False Positive Rate (FPR), False Negative Rate (FNR), Precision, Recall and F1-score (or F-score). We use the standard convention that positive samples are images that show a defect, and negative samples are defect-free images.

FPR and FNR were chosen because they are very meaningful statistics in defect detection, as they can be related to the potential associated industrial costs. The precision, recall and F-score are traditionally used in machine learning to compare algorithms when dealing with unbalanced classes. Other statistics, such as the ROC Area Under Curve or the Average Precision are measures of a family of models: they describe the properties of a set of models, usually defined by varying the decision threshold. These statistics have not been selected because they do not relate to a finalized decision system: as long as a decision threshold is not chosen, the model is not completely defined. For each class, the statistics are computed as follows:

- False Positive Rate: $FPR = FP/(FP + TN)$ (1)
- False Negative Rate: $FNR = FN/(FN + TP)$ (2)
- Precision: $Pr = TP/(TP + FP)$ (3)
- Recall: $Re = TP/(TP + FN) = 1 - FNR$ (4)
- F-score: $2Pr\,Re/(Pr + Re) = 2TP/(2TP + FP + FN)$ (5)

When the confusion matrix is not given in the original paper, N_{defect} and N_{clean} being respectively the number of images with and without defect in the test set, the entries of the confusion matrix are computed as follows:

- True Positives: $TP = Re\,N_{defect}$ (6)
- False Negatives: $FN = N_{defect} - TP = FNR\,N_{defect}$ (7)
- False Positives: $FP = TP(1 - Pr)/Pr = FPR\,N_{clean}$ (8)
- True Negatives: $TN = N_{clean} - FP$ (9)

The aforementioned performance statistics apply to image-level classification, without defect localization. We categorize these classification problems by their supervision level during learning. Here, we adopt the nomenclature of industrial defect detection, which differs from that of machine learning regarding "supervision". The categories of supervision are defined as follows:

- **Full supervision** refers to the use of all available features during learning, including the defect locations indicated by ellipses. For example, sliding window methods require a more or less precise location of the defect within the image.
- **Mixed supervision** refers to the use of all image-level labels (defective / non-defective) during learning as well as a fraction of the precise defect locations.
- **Weak supervision** refers to the sole use of all image-level labels during learning.
- **No supervision** refers to the absence of images of defect during learning.

We surveyed the evaluation protocols described in the various published articles to ensure a fair comparison of results. It is necessary to use a training/testing split of the data to ensure that the performance measures are unbiased. The DAGM dataset gives a default training/testing data split for each class, the use of which ensures comparable performance measures. However, these testing sets are rarely used in the literature: we have included all articles mentioning a training/testing split, even if this is not the default split.

We initially selected 60 articles by searching for articles using the DAGM dataset, using search engines like ScienceDirect and Google Scholar, or by using the references of other articles. Among these 60 articles, only 22 used test metrics compatible with ours. Indeed, some articles focus on segmentation performance only; others divert the DAGM dataset to classify the 10 textures; finally, some articles do not report statistics that can be related to FPR and FNR. Of these 22 articles, five do not describe a clear evaluation procedure, which compromises a fair comparison with other results.

Table 1 displays the performance statistics for the final selection of 17 articles. All results reported here are obtained from algorithms that have been trained and tested on at least two classes of the DAGM dataset. The numbers reported here are either extracted from the original article or calculated using explicitly available information. This can be the number of positive / negative samples used for testing, a confusion matrix, or simply a sub-part of the sought statistics.

Table 1. Performance statistics (in %) of different methods for industrial defect detection. Missing figures (not reported and not computable from given information) are marked with –. References marked in bold indicate that the source codes have been published by the authors. Classes refer to the number of datasets used for evaluation in each article.

Reference	Method	Classes	FPR	FNR	Precision	Recall	F-score
Unsupervised learning							
Zavrtanik 2021 [2]	AE	10	0.6	3.5	96.0	96.5	–
Tsai 2021 [3]	AE	2	9.5	5.5	90.9	94.5	92.7
Tsai 2021 (a) [4]	AE	2	7	2.75	93.3	97.3	95.2
Dekhtiar 2018 [5]	GAN	10	19.60	8.78	–	–	–
Weakly supervised learning							
Božič 2021 [6]	CNN	10	–	–	–	–	74.6
Lin 2020 [7]	CNN	10	0.07	0.55	–	–	–
Weimer 2016 [8]	CNN	6	0.75	0.95	–	–	–
Weimer 2015 [9]	CNN	6	4.66	2.58	–	–	–
Mixed-supervised learning							
Božič 2021 [6]	CNN	10	0	0	100	100	100
Fully supervised learning							
Lee 2020 [10]	CNN	6	0	0	100	100	100
Tabernik 2020 [11]	CNN	–	0	0	100	100	100
Kim 2020 [12]	CNN	6	0.06	0.3	–	–	–
Rački 2018 [13]	CNN	10	0.38	0.09	97.5	99.9	98.7
Wang 2018 [14]	CNN	6	0.05	1.13	99.6	98.9	99.2
Kim 2017 [15]	CNN	6	0.04	0	–	100	–
Scholz-Reiter 2012 [16]	Custom	5	1.74	1.78	–	–	–
Timm 2011 [17]	Custom	4	0.5	5.27	–	–	–
Siebel 2008 [18]	Custom	4	8.5	3.7	–	–	–

The references in bold indicate the articles for which the source code of the algorithms is published. Figure 2 provides another representation, by displaying the expected error rates for each algorithm, color coded to represent their supervision level.

Table 1 shows that the DAGM benchmark is solved for fully supervised approaches, in particular by the model of Božič et al. 2021 [6], which uses a CNN with segmentation and classification sub-networks. However, the dataset is still challenging in the weakly supervised and unsupervised frameworks. The weakly supervised method with the smallest FPR and FNR is Lin et al. [7] who introduce a novel CNN architecture trained using weak labels. Among the four unsupervised methods considered in this article, the method with the smallest FPR and FNR is Zavrtanik et al.'s DRAEM [2], which is an autoencoder (AE) trained using only negative, defect-free samples, and custom generated defects. The model still manages to output a full segmentation of the defect despite not using strong labels for training. It is also interesting to note that all unsupervised methods

use autoencoders or GANs, whereas weakly, mixed and fully supervised methods use CNN or custom algorithms. These custom algorithms are based on manually designed features (such as Local Binary Patterns) fed to a classifier, such as a Support Vector Machine (SVM) or a fully connected neural network.

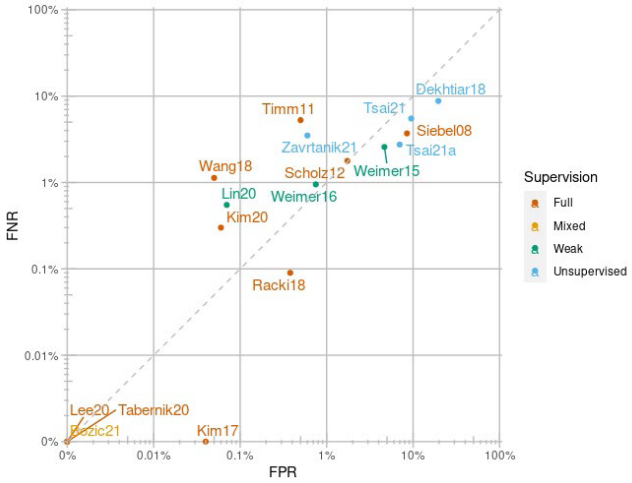

Figure 2. False negative rate versus false positive rate announced for the different algorithms (colour represents the level of supervision)

4. Discussion and conclusion

Out of an initial selection of 60 articles, we could only compare 17 of them because of evaluation procedure issues. On the one hand, some articles use dubious evaluation procedures, not clarifying how the algorithm was tested and trained. On the other hand, some articles do not use metrics that are relevant for a real industrial inspection system. We therefore advocate testing protocols using the default split between training and testing, as published by Wieler et al. [1], and reporting confusion matrices to improve the comparability of results. This would ensure the computation of all relevant statistics on the same basis, without bias in favour of either algorithm.

The results of Table 1 show that the DAGM dataset is solved by the best fully- and mixed-supervised methods. Many results can be considered to meet the industrial performance requirements, including those of some weakly-supervised methods. Industrial datasets are characterized by a strong imbalance of data in favour of non-defective items, a diverse defect typology, and variable environmental factors, such as temperature, vibration, or camera settings. Further testing on real-world datasets is necessary to confirm the applicability of the state-of-the-art models to real-world industrial inspection.

As expected, unsupervised methods, which use less expert information, are not as powerful as other methods. Figure 2 clearly shows that no unsupervised method reaches error rates in the order of the percent for both false positives and false negatives. Unsupervised methods can help answer industrial requirements, because they use no labelling for learning and they can adapt to new defect typologies. The use of these algorithms could therefore drive down the cost of production of datasets, and offer

increased performance for novel defect detection. We advocate developing and applying new unsupervised learning algorithms to industrial inspection datasets, including DAGM, with the aim to achieve performance suitable for industrial usage.

Acknowledgements

This work was carried out with the support of the Agence Nationale de la Recherche through the TEMIS ANR-20-CE10-0004 project.

References

[1] Wieler M, Hahn T. Weakly Supervised Learning for Industrial Optical Inspection; 2007. 29th Annual Symposium of the German Association for Pattern Recognition.
[2] Zavrtanik V, Kristan M, Skočaj D. DRAEM – A discriminatively trained reconstruction embedding for surface anomaly detection. In: International Conference on Computer Vision. IEEE; 2021. p. 8330-9.
[3] Tsai DM, Jen PH. Autoencoder-based anomaly detection for surface defect inspection. Advanced Engineering Informatics. 2021 4; 48.
[4] Tsai DM, Fan SKS, Chou YH. Auto-Annotated Deep Segmentation for Surface Defect Detection. IEEE Transactions on Instrumentation and Measurement. 2021; 70.
[5] Dekhtiar J. Deep Learning and Unsupervised Learning to automate visual inspection in the manufacturing industry. Université de technologie de Compiègne; 2019.
[6] Božič J, Tabernik D, Skočaj D. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. Computers in Industry. 2021 8; 129.
[7] Lin Z, Ye H, Zhan B, Huang X. An Efficient Network for Surface Defect Detection. Applied Sciences. 2020 9; 10(17).
[8] Weimer D, Scholz-Reiter B, Shpitalni M. Design of deep convolutional neural network architectures for automated feature extraction in industrial inspection. CIRP Annals. 2016 1; 65(1):417-20.
[9] Weimer D, Benggolo AY, Freitag M. Context-aware Deep Convolutional Neural Networks for Industrial Inspection; 2015. Workshop on Deep Learning and its Applications in Vision and Robotics, 28th Australasian Joint Conference on Artificial Intelligence.
[10] Lee H, Ryu K. Dual-Kernel-Based Aggregated Residual Network for Surface Defect Inspection in Injection Molding Processes. Applied Sciences. 2020 11; 10(22):8171.
[11] Tabernik D, Šela S, Skvarč J, Skočaj D. Segmentation-based deep-learning approach for surface-defect detection. Journal of Intelligent Manufacturing. 2020 3; 31(3):759-76.
[12] Kim S, Noh YK, Park FC. Efficient neural network compression via transfer learning for machine vision inspection. Neurocomputing. 2020 11; 413:294-304.
[13] Rački D, Tomaževič D, Skočaj D. A Compact Convolutional Neural Network for Textured Surface Anomaly Detection. In: Winter Conference on Applications of Computer Vision. IEEE; 2018. p. 1331-9.
[14] Wang T, Chen Y, Qiao M, Snoussi H. A fast and robust convolutional neural network-based defect detection model in product quality control. The International Journal of Advanced Manufacturing Technology. 2018 2; 94(9-12):3465-71.
[15] Kim S, Kim W, Noh YK, Park FC. Transfer learning for automated optical inspection. In: International Joint Conference on Neural Networks. IEEE; 2017. p. 2517-24.
[16] Scholz-Reiter B, Weimer D, Thamer H. Automated surface inspection of cold-formed micro-parts. CIRP Annals. 2012 1;61(1):531-4.
[17] Timm F, Barth E. Non-parametric texture defect detection using Weibull features. In: Fofi D, Bingham PR, editors. Image Processing: Machine Vision Applications IV. vol. 7877. International Society for Optics and Photonics. SPIE; 2011. p. 150-61.
[18] Siebel NT, Sommer G. Learning defect classifiers for visual inspection images by neuro-evolution using weakly labelled training data. In: Congress on Evolutionary Computation. IEEE; 2008. p. 3925-31.