Production-Ready End-to-End Visual Quality Inspection for Defect Detection on Surfaces Based on a Multi-Stage AI System

Patrick Trampert*[a], Tobias Masiak, Felix Schmidt[b], Nicolas Thewes, Tim Kruse, Christian Witte and Georg Schneider

Artificial Intelligence Lab, ZF Friedrichshafen AG, Scheer Tower II, Uni-Campus Nord, Geb. D5 2 66123 SB, Germany

Keywords: Machine Learning, Artificial Intelligence, Deep Learning, Visual Quality Inspection, Defect Detection, Windowing, Multistage Learning.

Abstract: Quality inspection based on optical systems is often limited by the ability of conventional image processing pipelines. Moreover, setting up such a system in production must be tailored towards specific tasks, which is a very tedious, time-consuming, and expensive work that is rarely transferable to different inspection problems. We present a configurable multi-stage system for Visual Quality Inspection (VQI) based on Artificial Intelligence (AI). In addition, we develop a divide-and-conquer strategy to break down complex tasks into sub-problems that are easy-to-handle with well-understood AI approaches. For data acquisition a human-machine-interface is implemented via a graphical user interface running at production side. Besides facilitated AI processing the evolved strategy leads to a knowledge digitalisation through sub-problem annotation that can be transferred to future use cases for defect detection on surfaces. We demonstrate the AI based quality inspection potential in a production use case, where we were able to reduce the false-error-rate from 16.83% to 2.80%, so that our AI workflow has already replaced the old system in a running production.

1 INTRODUCTION

The fast-developing ecosystem of the fourth industrial revolution with its ongoing digital transformation brings a lot of challenges and opportunities (Schwab, 2017). Hereby, intelligence-driven production is a central building block for success. Artificial Intelligence (AI) has empowered autonomous decision-makings in a production environment since the breakthrough of deep neural networks in 2012 (Krizhevsky et al., 2012). Production performance can benefit in multiple ways, for instance increasing transparency, higher efficiency, more flexibility, lightweight reconfiguration, easier controllability as well as cost optimization (Azamfirei et al., 2021a).

AI based production systems are subject to manufacturing prerequisites and have to fulfil many conditions to be integrated into an everyday shop floor. This starts at quality assurance regulations, includes cycle times and low scrap rates, and ends at employee empowerment. Important is not only a robust automation but also an integrated human machine interaction process to foster acceptance and scalability of AI. A tight intertwining of AI applications and shop floor workers ensures continuous monitoring and enhancement of implemented solutions to hold high quality standards, hence, a decrease of human imperfection (See et al., 2017).

Particularly in the context of Visual Quality Inspection (VQI) based on image processing, such a cooperation is extremely beneficial. An AI system can ask a human operator in case it is not confident about a decision. An initial uninformed AI will ask for help each time it gets presented an item to be analysed. With growing number of examples, the AI internalises human knowledge and needs less help. Furthermore, knowledge will be able transferable to similar problems.

Most VQI is based on specialised camera systems with built in or connected conventional image processing capabilities (Golnabi and Asadpour, 2007; Huangpeng et al., 2018). At Start of Production (SOP) for a new product great effort is needed to achieve full functioning of an image analysis system, for example to detect defects. Systems must be calibrated exactly for the new product or even for new

[a] https://orcid.org/0000-0002-1065-4345
[b] https://orcid.org/0009-0008-3010-8385
*Corresponding author

Trampert, P., Masiak, T., Schmidt, F., Thewes, N., Kruse, T., Witte, C. and Schneider, G.
Production-Ready End-to-End Visual Quality Inspection for Defect Detection on Surfaces Based on a Multi-Stage AI System.
DOI: 10.5220/0012536500003720

In *Proceedings of the 4th International Conference on Image Processing and Vision Engineering (IMPROVE 2024)*, pages 55-66
ISBN: 978-989-758-693-4; ISSN: 2795-4943

types or variants of already known products. In many cases, this is a tedious endeavour based on the knowledge and experience of the employee conducting this task. Invested time to guarantee a proper working of the system postpones the SOP and, hence, is very costly in terms of working time and lost production time. Technical changes in the production pipeline, like replacing a conventional oven for an induction oven, may require a recalibration as the altered system could fail to handle the resulting changes in error-patterns. Authorisation and approval granted to such a system each time involves many human experts, especially quality assurance personnel.

In this article we show how to replace this process with an AI based system for VQI of surfaces. The presented approach reduces calibration times to a minimum from several weeks to few hours, in particular for new types and variants of already known products, as the knowledge of associated AI models can be utilised. Our contributions are (i) introducing a workflow for learning an AI model for defect detection, which can assist from the start in examining produced parts by gathering and annotating data during production, (ii) AI supported interactive learning via Human Machine Collaboration (HMC) to digitalise process know-how, (iii) a flexible Graphical User Interface (GUI) for annotating images captured at production time as interface for the HMC, (iv) a multi-stage AI approach based on Deep Learning (DL) for optimised inference and cycle times, (v) cutting-edge AI technology combined with traditional manual design and modelling to have the advantages of both worlds while reducing the individual disadvantages, and (vi) end-to-end integration of the AI system into a shop floor.

Following, we conduct an exhaustive analysis of production research. We investigate the current state-of-the-art (SOTA) based on a use case explaining the motivation behind our work. This leads to the proposal of the developed workflow. The experimental setup as well as evaluation results follow afterwards. The article finishes with a discussion and conclusion of the presented work.

2 LITERATURE REVIEW

There is a wide range of applications, methods, and workflows for VQI at production lines. (Inman et al., 2003) performed a literature review on the intersection of quality and production system design and argue that the production system itself has a major impact on the quality. Hence, integrating an AI including a human machine interaction loop will improve quality. (Yazidi et al., 2018) designed and developed a distributed ontology-based quality control system and showed its feasibility for printed circuit boards. The approach enables autonomous quality control but lacks transferability to problems that are not covered by the ontology. (Raabe et al., 2017) proposed to introduce zero defect strategies or cyber physical systems to minimise scrap rates by use of self-correcting and autonomous systems. This reduces the dependence on manual decision-making processes and predictive condition-based maintenance to decrease human imperfection. An exhaustive review on the SOTA of quality control methods in the automotive manufacturing industry is presented in (Hafizi et al., 2019). Furthermore, the work of (Knop, 2020) gives very interesting insights into visual inspection methods and even investigates further concepts regarding the term “visual”.

Many examples of conventional image processing algorithms that aim to detect defects exist. (Tsai et al., 2012a) used an independent component analysis to identify errors in solar cells. This approach requires a lot of additional processing and relies on image reconstructions of defect-free images as baseline to classify unseen images. Besides big initial effort the approach is not flexible for similar problems and, hence, could only be evaluated on 80 test images resulting in a mean recognition rate of 93.4%, which is insufficient. Further approaches for solar panel defects, e.g., wavelet transforms (Li and Tsai, 2012) or Fourier transforms (Tsai et al., 2012b) could not increase performance significantly. (Chao and Tsai, 2010) proposed anisotropic diffusion to identify defects in low-contrast images. They applied their approach to material surfaces in liquid crystal displays. Such kind of model-based approaches require time-consuming adjustments per use-case. They are limited in terms of flexibility, computing demands, and applicability on production lines with short cycle times. (Jia et al., 2004) developed a system based on manually defined features as input for a support vector machine. Additional filtering is needed before using the system to find defects on surfaces of rolled steel. The performance and speed of the system made it real-time applicable. However, a huge effort was needed for implementation including manual modelling and human decisions intrinsic to the system. The presented approaches share several drawbacks. Such approaches often (i) only aim at a final step that makes a binary decision, which means it is not possible to extend the classification to more than two classes, (ii) require a huge effort to make them work, and (iii) are tailored towards one use case and application so that a transfer to other use cases is hardly possible.

Since the rise of DL in 2012 (Krizhevsky et al., 2012), many novel methods were investigated for the deployment to production. However, a lot of research remains on an experimental level or does not exploit its full potential. (Soukup and Huber-Mörk, 2014) used Convolutional Neural Networks (CNNs) to find rail surface defects. They used photometric dark-field stereo images. They showed that CNNs distinctly outperform model-based approaches, nevertheless, their setup is not suited to be used in day-to-day applications. (Du et al., 2019) used Faster R-CNN, which is an object detection approach, to detect casting aluminium defects based on X-ray images. They could increase accuracy in detection, yet the X-ray setup is quite expensive and very limited to specific applications. The lack of suitable datasets and the problem that new data is hard to obtain narrow down the usage further. A more elaborated example was shown by (Mueller et al., 2019b). They developed a binary real-time quality inspection powered by AI and validated the approach in an aerospace assembly to classify rivet connections during their joining process as OK or not OK (NOK). Besides the benefits, a disadvantage of their system is the limitation to only identify good and faulty parts, as it can't be extended to identify additional features, such as part numbers, engravings, or other features.

3 INTELLIGENCE-DRIVEN PRODUCTION

Nowadays, systems need high effort to be calibrated for a single use case for – at least partly – autonomous workflows. This involves a lot of tedious steps by-hand as well as testing to make sure that a single use case runs as it should (Malamas et al., 2003). Trial-and-error processes instead of an elaborated development for a defect detection pipeline are not rare. Quality control must ensure many sub-steps and regularly validate the system. In the worst case, this must be repeated every time something gets altered in the production (Kopardekar et al., 1993), for example if a new induction oven instead of a conventional is installed in a production line. A further problem is that many calibrated systems still perform quite average, so that additional manual re-inspections must be carried out. Even if the human effort at a production line is reduced by up to some percentage, the remaining percentage of re-inspection means additional work for unnecessary checking (Azamfirei et al., 2021b). An exhaustive review on the current SOTA in defect detection can be found in (Ren et al., 2021). The authors present a plethora of approaches that use machine learning. However, most of the presented methods are quite complex and require much time in advance to a possible usage. Besides all the semi-automatic workflows, there are even use cases in VQI that are completely manual human tasks without the help of cameras, or any autonomous system (Nessle Åsbrink, 2020).

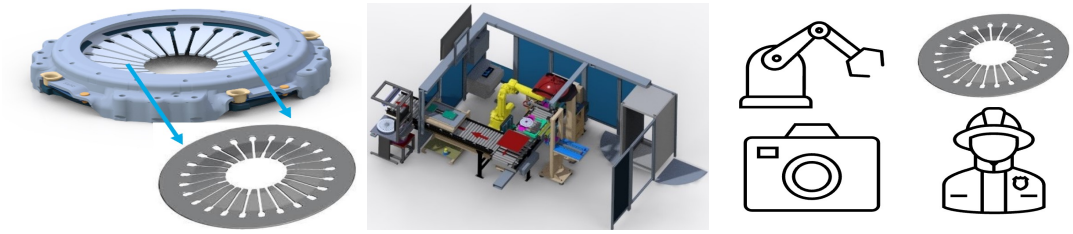
3.1 Current Production Setting

Based on a use case analysis and the planning approach "product – production process – production equipment" (Mueller et al., 2019a, cf.), the dependencies between a product, the current production or inspection process, and the current production equipment must be considered to derive a holistic solution concept (cf. Figure 1).

Example characteristics of a product can be shape, size, or material. Especially a multi-variant production as well as the production volume have an impact on the later process and, thus, on the selection of the production equipment needed for inspection. This use case is based on the manufacturing of a spring from the clutch pack of commercial vehicles. Springs are produced in many different variants. These variants can differ in terms of product size and the characteristics of various features, such as diameter, material thickness, or number of shaped holes.

The production process has constraints, e.g., cycle time. One process is an inspection task that analyses the surface of each spring to identify scratches, pressure marks, or material spelling and classifies these as NOK in comparison to OK parts with no faults (cf. Figure 2). In a non-digitalised environment, the inspection process was performed manually by humans in a monotonous and error-prone process. Since the advent of Industry 4.0, an increasing amount of inspections is handled by an automated cell. Such cells mostly use conventional image processing for workpiece classification.

The inspection cell consists of the following production equipment: A robot system for manipulating the components, a camera dome as well as a subsequent image processing system, and a human for re-inspection by-hand. The spring is delivered via a conveyor belt and automatically positioned under the camera dome with the robotic system. Several images are recorded from both sides of the component to capture the whole surface of each spring. The resulting images are prepared for further processing applying shape from shading (Zhang et al., 1999, cf.). This technique enhances small surface defects by exposing them from different angles with a flash, so that a three-dimensional impression of the surface is provided.



(a) In this use case a commercial vehicle spring is considered as a product.

(b) The production process is represented by an automated inspection at the production line.

(c) Used production equipment: robot, camera dome, conventional image processing, and human operators for re-inspection.

Figure 1: Planning approach for the development of an automated inspection process.

(a) pressure marks (b) scratches (c) material spelling (d) material inclusion

Figure 2: Examples of possible error features that need to be detected by the inspection system.

Image data is evaluated by a conventional image processing software and the classification result is reported to the cell controller. Depending on the result, in this case OK or NOK, the component is released for further assembly or made available to a human operator for re-inspection. The operator performs a final inspection of the spring via haptic feedback through palpating any scratches and grooves, or via visual feedback by means of a close investigation of the object. In case the component is defective, it is rejected and thrown away, otherwise it is released for further assembly. The current approach is far from being robust in terms of a high accuracy in part classification. Due to the complexity and individuality of possible defects, classical image processing has reached its limits. This results in a large number of parts incorrectly classified as NOK, also called false-NOK, which in turn leads to a time-consuming manual re-inspection.

The presented use case clearly shows the need for a more sophisticated workflow and enhanced algorithmic performance. Production should aim at a classification error rate converging towards zero eliminating many drawbacks mentioned before moving towards a fully automated production line. As a first step in this direction, we develop a workflow called Artificial Intelligence based Quality Inspection (AIQI). With AIQI we introduce AI to the production floor utilising new ideas of classical visual computing before AI processing to digitally optimise the data without the need of any optical requirements engineering in advance to image acquisition (Ren et al., 2021). Furthermore, we introduce a during production annotation via a GUI, which is developed from scratch and integrated into the production line. Combining SOTA AI techniques, we use the collected and annotated data to train a powerful and fast inferencing system that is deployed in production replacing outdated systems. As proof for the superior performance of our new workflow we conduct an evaluation of the old system against AIQI and show its significant impact on future production.

3.2 Material & Methods

AI systems usually need huge amounts of data to be trained from scratch, at the same time events of interest, here NOK-patterns, are seldom encountered and consequently under-represented in a collected dataset. To increase the amount of such data a technique called Data Augmentation (Wong et al., 2016; Xu et al., 2016) is used. Hereby, transformed versions of the original data are added to the training dataset. Common transformations from image processing can be used, like flipping, translations, zoom, lightning changes and many more.

DL for image processing mainly uses CNNs, which consist of many layers that can learn features, like low-level edges or mid-level characteristics of a provided problem (Yosinski et al., 2014). Based on this fact, such already trained features can be transferred to new, similar tasks, so that CNNs do not have to be trained from scratch, this process is called Transfer Learning (Shin et al., 2016). We identify three main CNN architectures through extensive ex-

periments to be used in the production line, namely ResNet (He et al., 2016), SqueezeNet (Iandola et al., 2016), and AlexNet (Krizhevsky et al., 2012), as they offer the best trade-off between performance and run time to guarantee short cycle times with high detection quality.

FastAI[1] serves as basis for our development (Howard and Gugger, 2020). This is a DL library that provides high-level components to easily use and apply AI techniques. Furthermore, low-level components are provided to add new functionality. *PyTorch*[2] (Paszke et al., 2019) serves as the AI backend framework to perform trainings. For being able to evaluate many different CNN architectures, we use the *Timm* library[3] (Wightman, 2019) as well as Torchvision[4]. These libraries offer a collection of image models and utilities to pull together a wide variety of SOTA architectures with the ability to reproduce ImageNet training results. All our trained models are exported to the exchange format Open Neural Network Exchange (ONNX) and executed in production with the runtime accelerator ONNX Runtime[5] to ensure fast inferencing times. Communication between the peripheries of the production station such as the Programmable Logic Controller (PLC) and our annotation tool or the inferencing application is realised with the integration of the multi-platform Ethernet communication suite Snap7[6]. The application and all its dependencies are bundled with the help of PyInstaller[7]. We bundle everything this way to guarantee isolation of system dependencies, and to ensure reproducibility.

3.3 AIQI Workflow

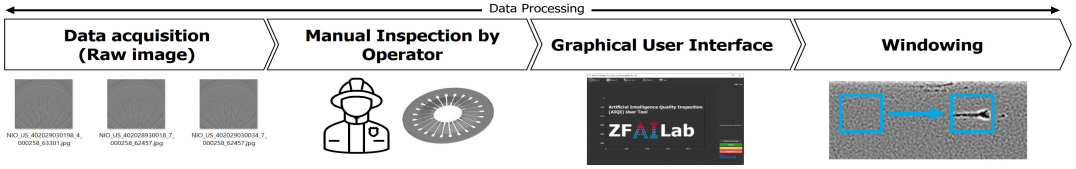
We present AIQI in this section. The main workflow consists of two pipelines, one for the collection of high-quality data in the production environment (Figure 3a and b) and one for the inference deployment at production lines (Figure 3c).

Before an AI can accurately classify the surface of, e.g., springs it is indispensable to acquire data and help the system to gather correct annotations. Human experts, like line operators or quality assurance staff, are involved in assisting an untrained AI to classify workpieces via inspection by-hand. Human involvement gradually decreases over time until an additional manual inspection is only required whenever new features arise, respectively when the AI is not confident about a classification. To accurately determine the confidence of an AI we also developed a method that enables the confidence estimation of an arbitrary CNN at run-time based on black-box access to the CNN. This method is used in AIQI, for further information have a look at (Woitschek and Schneider, 2022). The described data acquisition results in a data set that is put into the system by means of a HMC using an interface. This interface is tightly coupled within the production process via a GUI, developed especially for this purpose. Inspired by visual computing, images are split into small sub-images, also called patches, using a divide-and-conquer strategy adapted from sliding window approaches, called Windowing. First, the whole image of the workpiece is presented to a human operator. For critical surface defects the operator can select the image region that contains an anomaly. The resulting marked area is then split into patches, and these are presented to the operator besides the full view of the image. The operator must select and classify each patch as OK (marked green) or NOK (marked red) via a click operation in the GUI. The results of the human classification and the corresponding image data are used afterwards for AI model training. Each generated data point allows to improve the current classification accuracy via additional training of existing models.

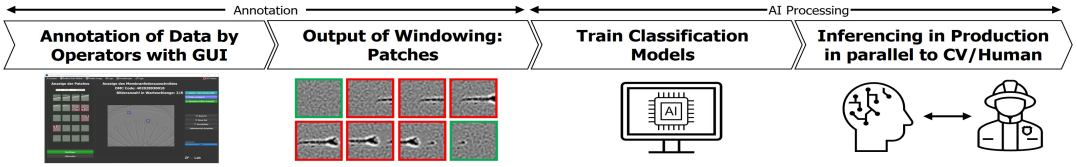
We introduce a visual computing step called Windowing (Schneider et al., 2023) as an additional stage in an AI based classification pipeline (see Figure 4). The grey coloured window at the top of Figure 4 a-c is moved with a defined step size over a defined region and determines sub-images of that region. Step size of the window in x- and y-axis direction as well as the size of the window $h \times w$ pixels can be modified at any time. However, the size of a single patch must be decided once and stays fixed. The objective of Windowing is to break down a large problem into many small entities as shown at the bottom of Figure 4a-c. Instead of searching for complex error features in high-resolution images, this approach allows the error features to be located and classified in very small sub-images. Thus, it is not necessary to specify each feature at a whole. It is sufficient to assign each sub-feature to its corresponding class.

The process therefore reduces the amount of data to be analysed in each case while keeping the original pixel resolution, which means there is no transformation of the data. The high-resolution sensor data is broken down into many small patches at original pixel resolution. In the workflow, as also shown in the use case, Windowing is applied to each large-scale image of the workpiece, so that many small patches,

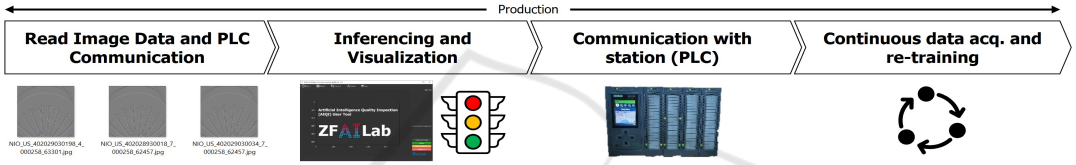
[1]https://github.com/fastai/fastai
[2]https://github.com/pytorch/pytorch
[3]https://github.com/rwightman/pytorch-image-models
[4]https://github.com/pytorch/vision
[5]https://github.com/microsoft/onnxruntime/
[6]http://snap7.sourceforge.net/
[7]https://pypi.org/project/pyinstaller/



(a) Data Processing in advance of applying our workflow is composed of data acquisition, manual inspection by human operators, setting up the annotation tool, and performing Windowing of the high-resolution large-scale images.



(b) Assisted by the implemented GUI, human operators annotate the sub-images coming from Windowing as OK or NOK. As soon as sufficient data has been collected, classification models can be trained and installed into production lines, where AI inference can run in parallel to the old computer vision (CV) system and human operators assist when AI lacks confidence.



(c) Each production line works independently now. Image data is acquired and read by the system via a PLC. The production version of our GUI takes care for AI processing, that means inferencing as well as visualisation. The PLC processes the results and saves them to a database. During this process additional data may be generated and used for continuous monitoring and training to enhance current models.

Figure 3: Artificial Intelligence based Quality Inspection (AIQI) Workflow.

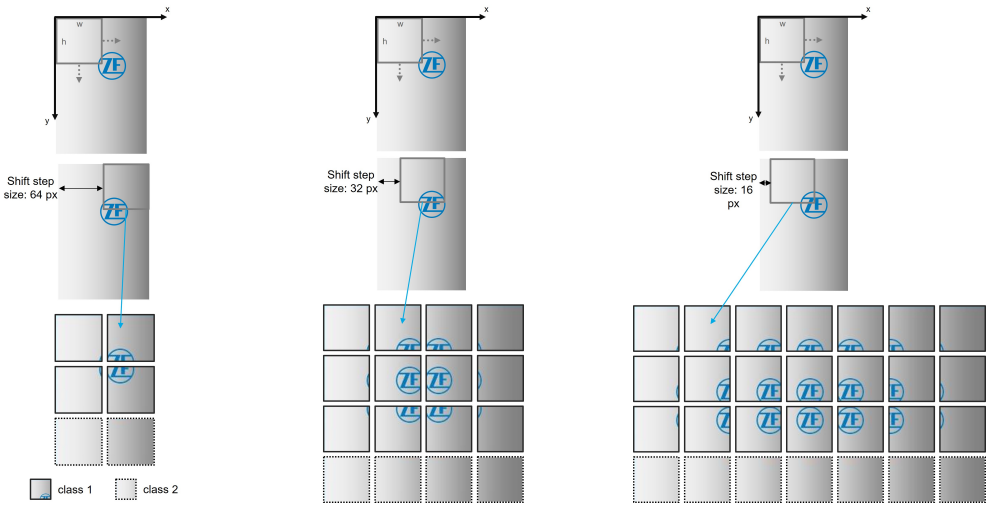
which are assigned to the OK or NOK class, are generated. Theoretically, Windowing is not only applicable to image data but can also be transferred to other sensor data.

Each individual patch is annotated and evaluated by-hand at the production start-up phase, as shown in Figure 4. The evaluation into class 1 and class 2, which means OK and NOK within the scope of quality inspection, is performed online during production via a GUI. The generated data and the associated classification labels are used for model training. During data annotation it is important to determine the ground truth based on the judgement of multiple human experts independently. Then, only data annotated consistently by everyone is kept ensuring increasing data quality for the ground truth as basis for AI training. At this point we want to emphasise once more the importance of data quality as already a small number of wrongly annotated data can disrupt AI model training significantly (Beggel et al., 2020, cf.).

In the presented workflow, a high quality of collected data and corresponding annotations must be ensured, which means each classification of each sub-image must represent the truth as defined by quality assurance. Each miss-classified data point will lead to a decrease in AI model performance. Hence, different procedures are integrated into the data quality process, which are applied to known annotated data as well as to unknown not yet annotated data. Examples are filtering based on high loss values both during training and validation, or confidence analyses of each pair of ground truth and predicted outcomes, for example each NOK that is really predicted as NOK.

As soon as a trained model is available, the classification of the patches is no longer performed by a human operator, but autonomously by the AI. In cases where the AI is unsure about the assignment as OK or NOK, e.g., because the error is a novel and very individual feature, the AI can ask the operator for advice and obtains help. The operator inputs the feedback via the GUI. The associated data is collected in the background and will be available for future model training. Through this approach, the accuracy of the model can be continuously improved.

Obviously, it does not suffice to classify only a single patch of several high-resolution images to assign the result with certainty to the workpiece. Hence, we developed a multi-stage inferencing pipeline for a



(a) (b) (c)

Figure 4: Divide-and-conquer approach via Windowing. (a) A window of size $h \times w$ pixels moves horizontally and vertically over a selected area with a configurable but fixed step size, which is also specified in pixels. Thereby, the image is split into many sub-images, also called patches. Patches are annotated by a human operator into classes, e.g., OK and NOK. Non-overlapping patches are generated in the initial Windowing step. (b) Stage two of our multi-stage classification of patches is shown. By altering the step size to a smaller value than the patch size, overlapping patches are generated. (c) The final stage with even smaller step size is shown. Combining these three, or even more, stages leads to enhanced feature detection.

reliable usage at production lines as depicted in Figure 4 as sketch. An initial inspection of each high-resolution image is performed by applying Windowing with maximal step size, so that non-overlapping patches are generated as shown in Figure 4a. The subsequent classification uses a fast AI model with a high false-NOK rate, which means many patches are identified as NOK although they are OK. The goal behind this procedure is to detect each possible defective patch, so that the likelihood to miss a NOK workpiece tends towards zero. All areas around identified NOK patches are analysed at least once more. These areas are sub-sampled with a smaller step size, so that overlapping patches are generated as shown in Figure 4b. Now, a slower more precise AI model is used to identify only patches that are really NOK, as it should be prevented to generate a lot of additional manual work due to unnecessary re-inspections. Depending on production constraints like cycle-times the procedure can be extended with a third stage as shown in Figure 4c. Subsequent stages are executed at different granularity applying AI models with varying accuracy towards less false-NOK predictions. The final patch classification can be further adjusted in terms of sensitivity. As example, a workpiece can be categorised as NOK based on a defined minimum of NOK patches, which could be one for a very high sensitivity and possibly more false-NOKs or five within a defined area size for less sensitivity. This number can be calibrated according to different workpieces and situations. The multi-stage procedure allows features to be inspected from different perspectives and provides a more accurate classification result adaptable to the needs of individual use cases. At this point AIQI benefits from the usage of small patches in contrast to whole high-resolution images because novelties can be analysed completely isolated from workpiece geometries and image sizes The AI is not confused and mislead from the surrounding environment near any anomaly, which often is a problem when applying DL (Heaven et al., 2019).

A running production line incorporates additional modules (cf. Figure 3c). An additional production-ready version of our GUI is installed and serves as brain of the quality inspection. Besides running the AI inference, results are visualised and the HMC takes place when the AI asks for it. Furthermore, results are processed and saved to a database via a PLC. Additional data may be generated at this point for continuous monitoring or to enhance current models.

4 USE CASE RESULTS

We show the usefulness and impact of the developed workflow by means of the presented workpiece production of springs. Two production lines were selected as basis for the AIQI development and evaluation. Besides the already described workflow, we

used SOTA techniques from AI research and development to ensure that no potential is wasted. Before presenting the setup and results the most important ingredients will be recapped.

We selected *PyTorch* and *FastAI* as main development frameworks for implementation as well as training and evaluation of our AIQI workflow. Due to low data availability and faster convergence rates all models were learned with pre-trained architectures.

As already shown before, cf. Figure 3, we had to start with data acquisition as not a single image was available at that time. The whole pipelines including annotation routines, image data quality assurance, and the GUI were developed upfront and set up in production. During data collection we developed the training and inferencing routines and tested these regularly with increasing data availability to ensure robust working of the software.

For the experiments we used 19,265 patches partitioned into the classes OK and NOK with 15,600 OK and 3,665 NOK. As AIQI is able to include an arbitrary number of classes with no additional effort, we tweaked the training utilizing this fact. OK patches were split into the classes OK and false-NOK, which resulted in 12,638 OK and 2,962 false-NOK. The AI was then trained with three classes, and a subsequent inference was mapped onto the classes OK and NOK again for decisions on large-scale images in production. The applied trick halved the error-rate in upcoming results. Due to production constraints, in particular cycle times, not only the best performance in terms of accuracy was important, but also a runtime smaller than allowed. Hence, we evaluated 1065 different CNN architectures to identify an optimal trade-off. We refrain from presenting all of these results, as this would not add any value to the presented work. The data patches were split into a training and a validation set with an 80 to 20 ratio and equally distributed classes in both sets. For testing we gathered additional data sets of workpieces that were completely independent of the training patch data set. The whole inferencing workflow was applied to all high-resolution images.

Based on the experiments, we identified three architectures for the final inferencing pipeline used in the described multi-stage approach. The multi-stage approach was fixed using SqueezeNet in round one, AlexNet in round two, and ResNet as final decision step for the classification. This results in a very fast first stage and a very accurate last stage. We only report performance in terms of accuracy for ResNet, as stage three determines the final classification. For the training set an error rate of 0.0325% was achieved, which shows that training converged quite well. The validation set delivered an error rate of 3.5%, which was the best possible trade-off between accuracy and run-time also incorporating the additional stages as pre-selection and acceleration steps.

Before installing the AI in production as independent system, we had to convince quality assurance that our workflow outperforms the old one. Hence, a quality assurance approved data set of 386 whole images, which consisted of 298 OK and 88 NOK workpieces, was composed and approved for a final test. The old system was compared against AIQI. An important remark here is, that both systems did not confuse real NOK workpieces as OK, so that no defective parts passed the quality check. Hence, the goal was to reduce the false-NOK rate while keeping the false-OK rate at zero. The old system had an error rate of 27%, whereas our AI workflow had an error rate of 3.1%. This convinced quality assurance to install the AI in parallel to the old system for a long-term test. The following months a long-term test was performed running both systems in parallel. The long-term evaluation was based on 9,020 workpieces. The old system performed at an error-rate of 16.83% and the AI at 2.80%. In sum, the validation set, the test set, as well as the long-term test set performed consistently in terms of error rates, which is a strong hint for a well-trained AI. Consequently, the old system was shut down after the evaluation.

5 DISCUSSION

After having presented AIQI and an use case application that is already deployed at a production floor including superior results compared to a traditional workflow (cf. Figure 5), we concentrate on the individual steps and work out the benefits and limitations of AIQI.

Compared to more complex AI methods, e.g., bounding box annotation in object detection, initial data collection and the connected annotation workflow during production are much easier with AIQI. Firstly, due to system design as annotation and production are not separated processes, secondly, as the annotation workflow itself exhibits a strongly reduced complexity. Looking at bounding boxes, tagging features does not lead to unique results for different annotators as the variability in bounding box sizes needs much more tedious work and, hence, is more error prone despite a higher effort. In our approach there is a clearly defined workflow to generate distinct sub-images and it is only needed to assign them to a class. Furthermore, the reduced complexity makes our approach less data hungry than standard bounding box

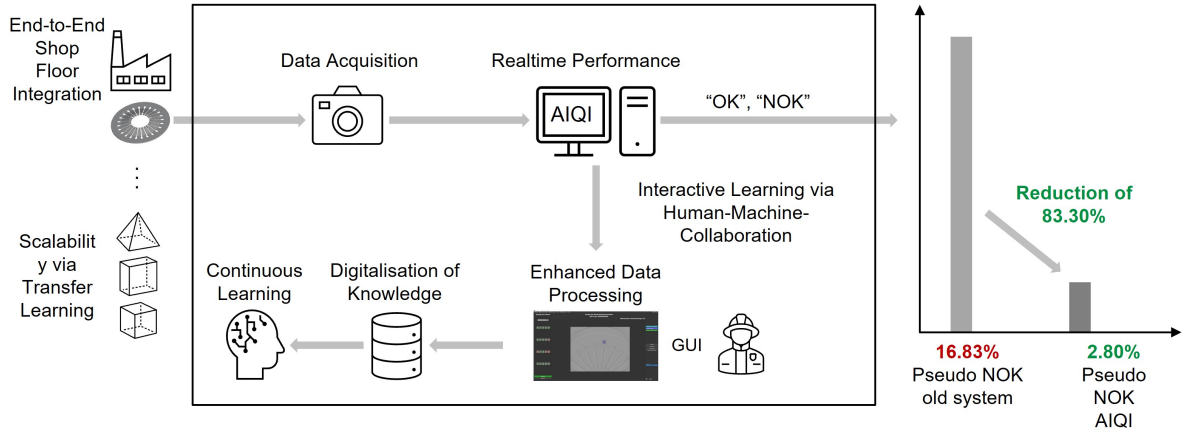


Figure 5: Summary of our contributions and achievements as workflow for AI based defect detection. Enhanced data processing via Windowing is one of the crucial steps within a shop floor integration to deliver an End-to-End solution. Acquired sensor data is analysed in real time by AIQI utilising its multi-stage AI approach based on DL. Cutting-edge AI technologies in combination with conventional methods integrate the advantages of both worlds. Hereby, interactive learning is realised via HMC to digitalise process know-how by means of a GUI and its annotation functions. The system is easily scalable because of integrated transfer learning methods toward further applications. In the presented use case, AIQI reduced the pseudo error rate, depicted as Pseudo NOK, of the old system by 83.30% from 16.83% down to 2.80%.

approaches. Other ways of defect identification are at least as complex as object detection which highlights the importance of our AIQI approach.

The proposed annotation workflow developed in our GUI profits from an effect called Gaming With A Purpose (GWAP) as described in (Venhuizen et al., 2013). Hereby, labelling tasks are presented as games to the user, but serve a purpose to the owner of the product, respectively the principal of the annotation task. Such GWAP's are nowadays used in the form of captchas where each user identifies objects by clicking on boxes. Compared to bounding box approaches, where it is needed to draw appropriate boxes for defects, clicking on readily prepared boxes is much easier. This increases the motivation of the user for the data labelling task to do a good job. The authors have shown that an increased motivation results in higher quality when annotating data, which was one more reason we developed AIQI. We are aware that the facilitated annotation workflow does not solve each kind of problem. Humans still suffer from imperfection, and it is quite likely that errors occur during labelling that can mislead any AI training. Hence, ground truth data must be independently validated by different persons and the current workflow must be challenged and developed further. AI-based support via trained models as well as semi-automated labelling assistance can also be enhanced further to reduce labelling errors and effort.

Besides the advantages for data labelling, Windowing also brings significant algorithmic advantages. Identifying defects and anomalies on large-scale images without losing pixel resolution and, hence, detection quality is very challenging. Analysing whole images as basis for classification is very compute intensive and requires a lot of data covering the possible space of occurring situations like specific errors. Applying Windowing, which is a divide-and-conquer inspired splitting of a large-scale image into smaller sub-images, resolves such issues and facilitates many tasks. Hereby, a large complex problem is broken down into many small and simple problems that can be classified more easily and efficiently. Furthermore, training an AI with patches that are context independent of its environment prevents focusing on the surroundings, which increases the ability to learn what is expected.

When cutting an image into many small pieces to find erroneous spots it is important to create some logic around the whole system arising from this idea. Classification of a single patch isn't enough to determine the quality of a produced part robustly. Consequently, we developed a multi-stage approach, which has the advantage to be more reliable due to multiple factors used for a decision, so that more knowledge than only single patch classifications is incorporated. The multi-stage approach also has a higher degree of flexibility to adapt to the classification process. For example, sensitivity can be varied in subsequent steps by refining the minimum error size or the number of patches with detected errors that are needed to judge a whole image as containing at least one defect or anomaly. Nevertheless, a multi-stage AI is often not usable out of the box, hence, additional work may be

needed to modify AIQI.

Windowing is tailored towards basic classification tasks, which are well researched, fast, robust, and do not require complicated workflows or architectures. In a production setup, image processing speed is one of the main bottlenecks that influences real-time performance of vision systems (Ren et al., 2021). The simplicity of our approach to break down a hard task into many easy tasks delivers a huge benefit towards this problem, as we showed in the depicted use case. It may be misleading that we argue with shorter runtimes and easier application, as there are additional steps involved that cost time and need to be integrated. However, the developed multi-stage pipeline makes it possible to identify important areas of large-scale images to only concentrate on small percentages for the complex operations. The need for such an initial search may require some time but saves much more time in subsequent steps. Furthermore, we sped up inferencing times by converting models to the exchange format ONNX. Utilizing the ONNX format and its runtime environment ONNX Runtime provides fast computation times, so that the inferencing time can be sped up drastically compared with data formats from other frameworks, like *PyTorch* or *TensorFlow*. This is crucial in time sensitive applications, especially on the production floor, where each part has a predetermined cycle time for each step to minimise delays and maximise the number of produced parts.

Additional benefits emerge that make future developments easier and more reliable in terms of tackling new use cases. Patches can be collected as kind of an error catalogue for surface defects and anomalies. This offers an adaption onto other variants from a workpiece, like springs with a varying diameter or differing holes, out of the box with minimal additional effort. Furthermore, other produced workpieces also share properties regarding their surface, like the materials they consist of, hence, also errors occurring on such surfaces look similar. Consequentially, a knowledge transfer from collected and annotated data of workpieces to new or not yet included workpieces is possible. Due to the splitting into small sub-images the collected error catalogue becomes even more informative. Instead of templates for complete errors, we collect subsets or fragments from varying angles and positions. Using such sub-error parts or sub-anomaly parts the developed multi-stage approach makes it possible to put these together like in a jigsaw and recognise even unseen complete errors that consist of the sub-parts. That means, scalability becomes simpler, and it is like working on the same data as size of images and defects does not play an important role for sub-images of the same small size. With a growing number of use cases this leads to a comprehensive error database for surface defects and anomalies as a digitalization of human knowledge that can be utilised for the development of powerful AI methods.

Where there is light, there is also shadow. We should not expect to have solved all problems regarding the detection of surface errors and anomalies as soon as we identified a use case. There will always occur defects never seen before, or materials that look the same when OK may develop different properties and challenge the AI based quality inspection. Close monitoring and regular human quality controls still need to be performed besides the ability of the AI to ask for help and additional data that can be added to the models via continuous learning strategies. Exactly such strategies were put into practice in the spring use case during the last year. Additional data was collected and annotated based on uncertainties from AI as well as from observed problems for continuously updating running models. Based on the most recently produced 16,153 workpieces we performed a long-term comparison of the initially deployed and latest up-to date models. The initial setup led to an error rate of 4.42%, which shows that over time models can decrease as well as that not all kind of errors had been covered. The latest setup led to an error rate of 2.58%, which is even better than the final evaluation.

Summarizing based on the discussed advantages and limitations, AIQI turned out to be a promising addition to a digitalised production incorporating computer vision and machine learning. Realising more use cases and utilising the capabilities already present will give additional insights to further extensions and a rich database for surface defect recognition.

6 CONCLUSION

The main contribution of our work is a multi-stage AI system for image processing tasks. Instead of using a sophisticated black box like end-to-end algorithm that is not configurable in between, we propose to breakdown a complex task into small and easy sub-tasks that can be handled with well-understood basic AI approaches. Besides enhanced control over data collection and training, AIQI is more flexible due to manual modelling and configurability. This makes it possible to introduce additional steps based on, e.g., defined metrics and confidences to control the behaviour of the whole process. Furthermore, it is easier to understand the inner workings and, hence, to explain the system to operators that use it in day-to-day production. The ability to explain, what the AI is doing and why, increases transparency and hence acceptance of

the system. The interwoven human machine collaboration ensures both digitalisation of human knowledge and high-quality performance due to continuously updated data and trained models. The AI system asks for help if unclear instances are detected and incorporates the human answers into future assessments.

The presented workflow can be extended in several ways, like introducing a higher degree of detail than only OK and NOK labels as we did by introducing false-NOK labels, adding more stages if needed, changing the basic AI algorithm, adapting existing models to similar problems via transfer learning, or tuning sensitivity based on various metrics and defined confidence measures. Numerous options for manual modelling combined with SOTA AI algorithms make the process versatile. A further option is the generation and use of synthetic data, which could be incorporated into future models. Such data could even enforce a deployment at production start without ever having produced any part before.

Scalable and accurate quality inspection with little maintenance effort is fundamental for production. The implemented workflow and all the automated processes as well as the easy deployment as application make our AI-based quality inspection workflow a great tool for a fast time-to-market due to short and efficient development cycles. The knowledge digitalisation of human experts offers a huge business value. There is no need to artificially define any error catalogue of what could happen but rather it suffices to collect and annotate data based on a running production line. The resulting knowledge database is not only more precise than predicted anomalies for a workpiece but also is transferable to all kind of similar problems in the future.

We conclude that our workflow is suited for a wide range of applications and, hence, should be taken into consideration for VQI tasks in a production setting.

ACKNOWLEDGEMENTS

We thank the shop floor in Schweinfurt, especially Michael Huebner and Daniel Joerg for the close collaboration and their continuous availability. Furthermore, we thank Andreas Dorsch who made this project possible with his enthusiasm and ongoing support.

7 FUNDING

The research was supported by funding from Bundesministerium für Wirtschaft und Klimaschutz (BMWK) in the project Vernetzter digitaler Assistent für das Datengetriebene Engineering von Roboterbasierten Produktionsanlagen (VADER, 01.01.2023 – 31.12.2025).

REFERENCES

Azamfirei, V., Granlund, A., and Lagrosen, Y. (2021a). Multi-layer quality inspection system framework for industry 4.0. *International Journal of Automation Technology*, 15(5):641–650.

Azamfirei, V., Granlund, A., and Lagrosen, Y. (2021b). Multi-layer quality inspection system framework for industry 4.0. *International journal of automation technology*, 15(5):641–650.

Beggel, L., Pfeiffer, M., and Bischl, B. (2020). Robust anomaly detection in images using adversarial autoencoders. In Brefeld, U., Fromont, E., Hotho, A., Knobbe, A., Maathuis, M., and Robardet, C., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 206–222, Cham. Springer International Publishing.

Chao, S.-M. and Tsai, D.-M. (2010). Anisotropic diffusion with generalized diffusion coefficient function for defect detection in low-contrast surface images. *Pattern Recognition*, 43(5):1917–1931.

Du, W., Shen, H., Fu, J., Zhang, G., and He, Q. (2019). Approaches for improvement of the x-ray image defect detection of automobile casting aluminum parts based on deep learning. *NDT & E International*, 107:102144.

Golnabi, H. and Asadpour, A. (2007). Design and application of industrial machine vision systems. *Robotics and Computer-Integrated Manufacturing*, 23(6):630–637.

Hafizi, M., Jamaludin, S., and Shamil, A. (2019). State of the art review of quality control method in automotive manufacturing industry. In *IOP Conference Series: Materials Science and Engineering*, volume 530, page 012034. IOP Publishing.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Heaven, D. et al. (2019). Why deep-learning ais are so easy to fool. *Nature*, 574(7777):163–166.

Howard, J. and Gugger, S. (2020). Fastai: A layered api for deep learning. *Information*, 11(2).

Huangpeng, Q., Zhang, H., Zeng, X., and Huang, W. (2018). Automatic visual defect detection using texture prior and low-rank representation. *IEEE Access*, 6:37965–37976.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). Squeezenet:

Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Inman, R. R., Blumenfeld, D. E., Huang, N., and Li, J. (2003). Designing production systems for quality: research opportunities from an automotive industry perspective. *International journal of production research*, 41(9):1953–1971.

Jia, H., Murphey, Y. L., Shi, J., and Chang, T.-S. (2004). An intelligent real-time vision system for surface defect detection. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 239–242. IEEE.

Knop, K. (2020). Indicating and analysis the interrelation between terms-visual: management, control, inspection and testing. *Production Engineering Archives*, 26.

Kopardekar, P., Mital, A., and Anand, S. (1993). Manual, hybrid and automated inspection literature and current research. *Integrated Manufacturing Systems*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105.

Li, W.-C. and Tsai, D.-M. (2012). Wavelet-based defect detection in solar wafer images with inhomogeneous texture. *Pattern Recognition*, 45(2):742–756.

Malamas, E. N., Petrakis, E. G., Zervakis, M., Petit, L., and Legat, J.-D. (2003). A survey on industrial vision systems, applications and tools. *Image and vision computing*, 21(2):171–188.

Mueller, R., Franke, J., Henrich, D., Kuhlenkoetter, B., Raatz, A., and Verl, A. (2019a). *Handbuch Mensch-Roboter-Kollaboration*. Carl Hanser.

Mueller, R., Vette, M., Masiak, T., Duppe, B., and Schulz, A. (2019b). Intelligent real time inspection of rivet quality supported by human-robot-collaboration. *SAE Technical Paper*, 2(2019-01-1886).

Nessle Åsbrink, M. (2020). A case study of how industry 4.0 will impact on a manual assembly process in an existing production system: Interpretation, enablers and benefits.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Raabe, H., Myklebust, O., and Eleftheriadis, R. (2017). Vision based quality control and maintenance in high volume production by use of zero defect strategies. In *International Workshop of Advanced Manufacturing and Automation*, pages 405–412. Springer.

Ren, Z., Fang, F., Yan, N., and Wu, Y. (2021). State of the art in defect detection based on machine vision. *International Journal of Precision Engineering and Manufacturing-Green Technology*, pages 1–31.

Schneider, G., Masiak, T., Trampert, P., and Schmidt, F. (2023). Prüfen eines prüflings, patent number 10 2021 210 572.6.

Schwab, K. (2017). *The fourth industrial revolution*. Currency.

See, J. E., Drury, C. G., Speed, A., Williams, A., and Khalandi, N. (2017). The role of visual inspection in the 21st century. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, pages 262–266. SAGE Publications Sage CA: Los Angeles, CA.

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., and Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298.

Soukup, D. and Huber-Mörk, R. (2014). Convolutional neural networks for steel surface defect detection from photometric stereo images. In *International Symposium on Visual Computing*, pages 668–677. Springer.

Tsai, D.-M., Wu, S.-C., and Chiu, W.-Y. (2012a). Defect detection in solar modules using ica basis images. *IEEE Transactions on Industrial Informatics*, 9(1):122–131.

Tsai, D.-M., Wu, S.-C., and Li, W.-C. (2012b). Defect detection of solar cells in electroluminescence images using fourier image reconstruction. *Solar Energy Materials and Solar Cells*, 99:250–262.

Venhuizen, N., Evang, K., Basile, V., and Bos, J. (2013). Gamification for word sense labeling. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*.

Wightman, R. (2019). Pytorch image models. https://github.com/rwightman/pytorch-image-models.

Woitschek, F. and Schneider, G. (2022). Online black-box confidence estimation of deep neural networks. In *33rd IEEE Intelligent Vehicles Symposium (IV22)*.

Wong, S. C., Gatt, A., Stamatescu, V., and McDonnell, M. D. (2016). Understanding data augmentation for classification: when to warp? In *2016 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–6. IEEE.

Xu, Y., Jia, R., Mou, L., Li, G., Chen, Y., Lu, Y., and Jin, Z. (2016). Improved relation classification by deep recurrent neural networks with data augmentation. *arXiv preprint arXiv:1601.03651*.

Yazidi, K., Darmoul, S., and Hajri-Gabouj, S. (2018). Intelligent product quality control and defect detection: A case study. In *2018 International Conference on Advanced Systems and Electric Technologies (IC_ASET)*, pages 98–103. IEEE.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

Zhang, R., Tsai, P.-S., Cryer, J. E., and Shah, M. (1999). Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706.