Challenges in evaluating drone data

Words: DI Christian Raml, Head of R&D at Aero Enterprise GmbH

Nowadays there is apparently almost no problem that cannot be solved with AI. If cars can even drive completely autonomously, should it not be easy then to find defects and other abnormalities on images from rotor blades of wind-turbines? However, where is the algorithm that detects defects in real time, qualifies them and gives a recommendation for action?

Compared to engineering autonomous vehicles, the problem sounds very simple. Still, it is much more complex than it looks at first glance. The topic of this article is about identifying the difficulties in this business field and what is necessary to overcome these obstacles.

Quality and quantity

Depending on the size, the average surface of the rotor blades on a wind turbine makes up 1000 to 2000 m², which must be inspected visually. In an automated flight, a drone covers this area within just 20 minutes. At Aero Enterprise, the surface is resolved with a median resolution of 20 px/mm². The resulting data record consists of hundreds of images, depending on the dimension of the turbine and the desired image overlap. In one day, for example, inspecting eight turbines with a mean rotor diameter of 135 meters would result in about 8000 pictures. That requires 480 GP (Giga pixels) to be processed and evaluated each day.

This introductory example shows an essential aspect: nowadays, the problem is no longer the quality of the data but it is the quantity. Processing such a volume by hand is almost impossible. That is where AI comes in quite handy.

Is it possible to extract the necessary information out of a given dataset without human interaction?

Digging a little bit deeper and trying to identify the steps and tools, which are necessary to create, develop or train a system that can solve the problem. At the beginning it is necessary to identify and define what we are interested in. The goal is to create high quality training data, which is one of the most important steps.

There are many theories about how perfect training data needs to look like. Two key properties are consistency and relevancy.

To learn from a dataset, it should contain representative examples. In the best case, humans would also be able to learn from the dataset. Of course, it would be very bad if the information is not consistent. For example, referring to a contour as a circle and in another picture to the same contour as a cuboid. Furthermore, reducing the information to the important and relevant points mean learning will be faster.

Learning from pioneers

A comparison between a similar, already well-solved problem can help to identify solutions. As an example, let us take traffic sign recognition into account, where areas of interest in images are classified, and which works excellently nowadays. The available amount of good training data is definitely a reason why this task works so well. Why is this the case?

During one's driving lessons, participants study a separate chapter that deals with traffic signs. Using the course materials, each participant learns what traffic signs look like and what their meanings are. There is no room for interpretation, everyone is able to detect and classify traffic signs on an image. This task also can be executed extremely fast because it is obvious where to look for them: along the road.

If we were to extend the task to also evaluate

Christian Raml

signposts, it gets more difficult. Signposts are no longer beside the road. They appear on posters in the distance or on buildings. Sometimes the information is not immediately visible because it is attached to an advertisement. People may interpret them in different ways as there are no requirements for signposts.

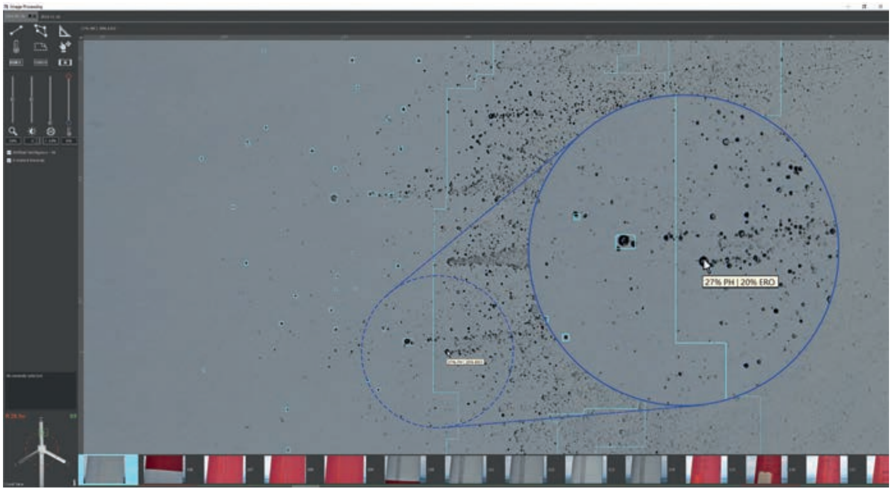
The human factor

In contrast to traffic signs, there is no standard for identifying and interpreting defects on rotor blades. Even experts sometimes disagree about defects. Of course, guidelines and recommendations exist, but the requirements differ by country, region, and company/customer. Defects on rotor blades are in some way equivalent to signposts. There is room for different interpretation. As one driver easily can follow the route description attached to an advertisement poster, the other one does not even recognize the ad as signpost.

If two rotor blade experts identified areas with defects, even assuming they used the same source of information, perhaps the outcome would not be exactly the same. If they both created training data for a neuronal network according to their conception, the output would then have some uncertainty.

Things get even more difficult if they also define the defect class for each area of interest. For example, in addition to the erosion class, some companies have another distinction between leading- and trailing-edge erosion. They can be treated as subclasses or individual classes. Pinholes, cracks, holes, burns by lightning strikes and erosion are always linked with damaged coating, which is in most cases a distinctive class too.

Classification is very subjective. For a group of pinholes, every pinhole can be marked as an individual abnormality or, in a small group, as one defect of class erosion. On the other



Anomaly interpreted with two types of defects

hand, creating rules to distinguish exactly between these two classes does not make sense, because the guidelines would be extraordinarily complex. We can try to use pinholes in a specific area as boundary. That would mean that we need to count the pinholes. Then we have to define the 'area': quadratic, rectangular, circular, or elliptical? Is the position also important? This example shows that defining exact rules for all types of defects is extremely hard.

Classification guidelines

As a consequence, each expert will evaluate according to his or her own experience and some 'soft' guidelines. If we use the information to train a neuronal network, then the output will be somewhere in the middle of all these different opinions. The training process is nothing more than creating rules to match all the views in the best feasible way. The important thing is that the output could be different to our own estimation. And if this is the case, the result of the AI is not necessarily wrong.

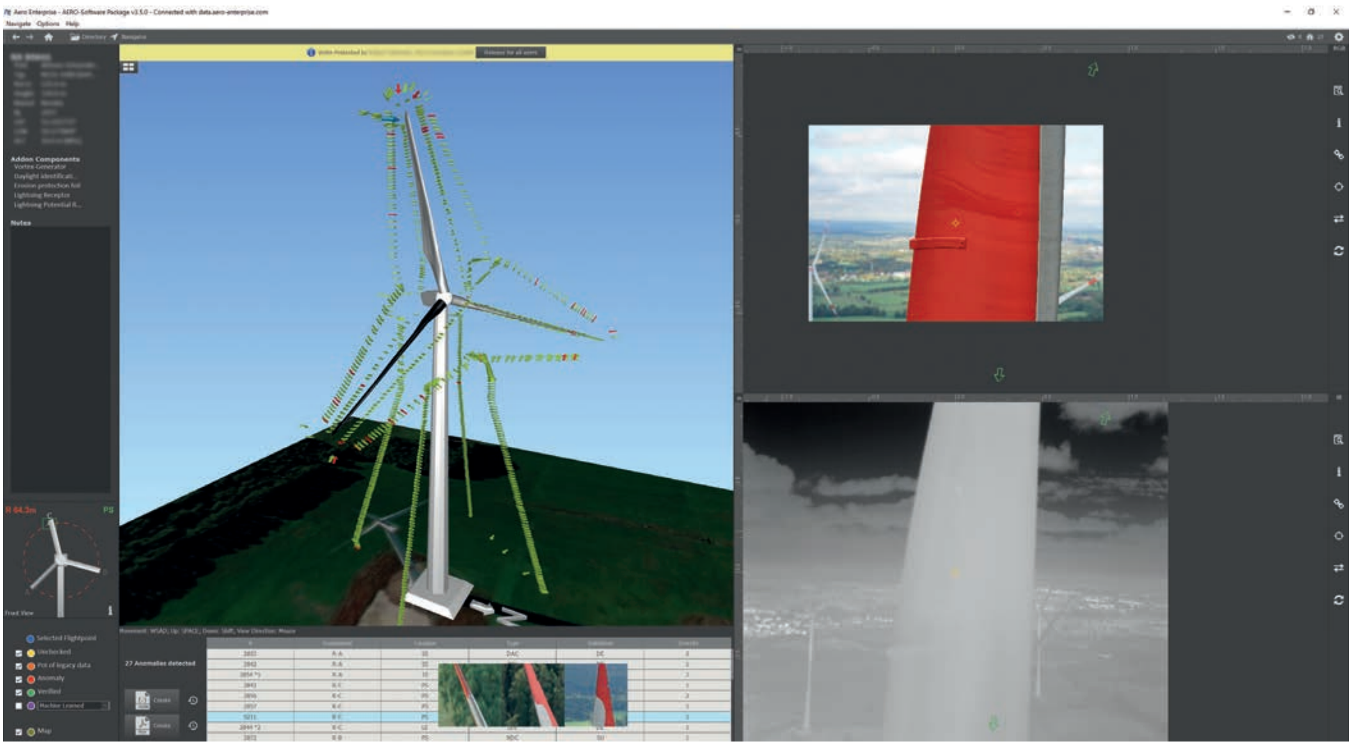
Referring to relevancy, is it necessary that an AI can differ between the same classes like a human? Does it even make sense to note pinholes?

To determine how relevant a defect is, further information and additional experience are usually required. This property is rated more or less with the severity grade. In theory a neuronal network learns this information indirectly. If every defect of a certain class is labelled on every image in the dataset, the trained network has a higher probability of identifying this class correctly, it is more relevant.

Now, the question may appear if there are defects which are not always labelled?

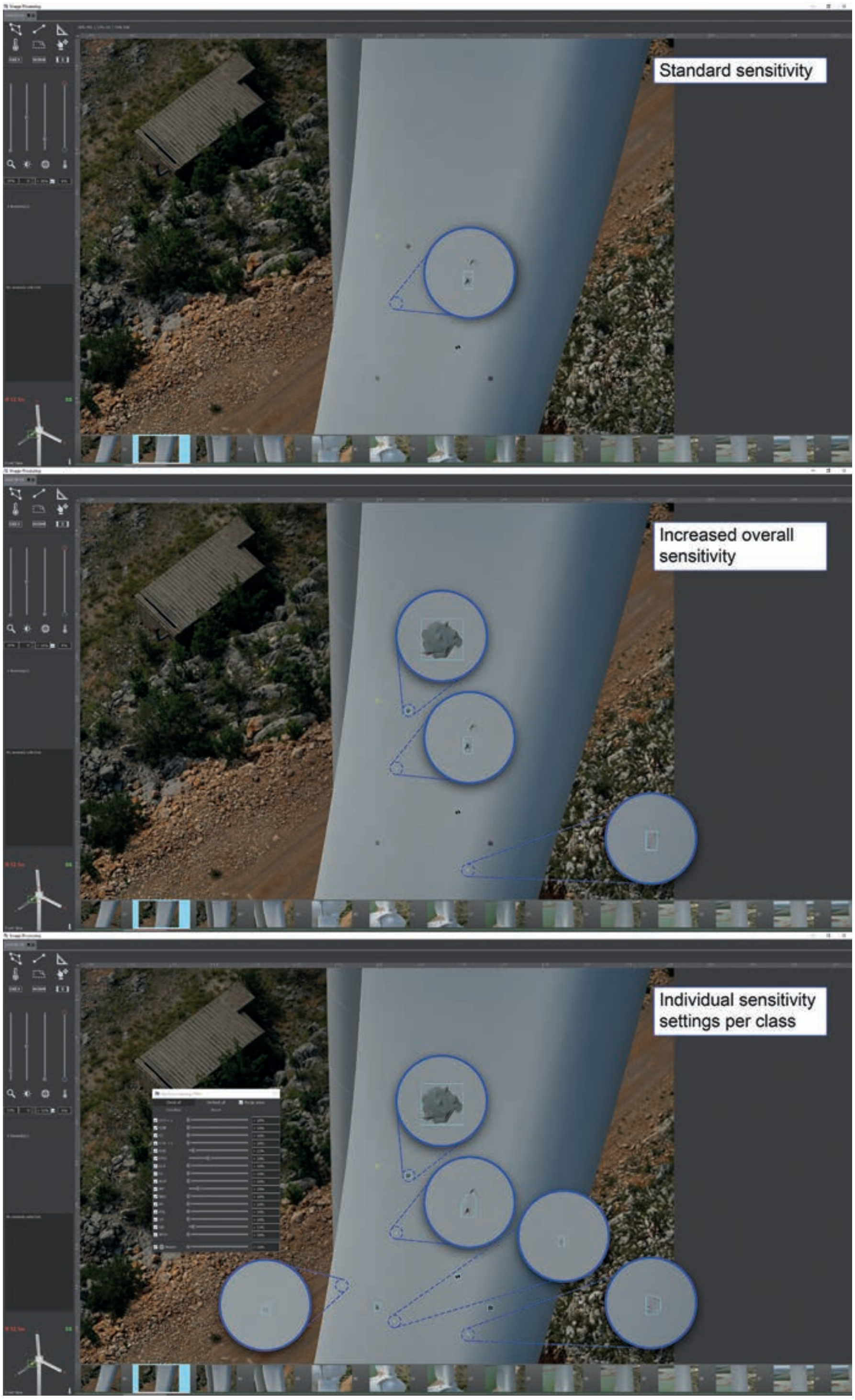
For pinholes this is the case. As stated previously, a group of pinholes is normally treated as erosion or damaged coating. Not every single pinhole is marked as an individual defect.

Alternatively, severity grade information can be provided directly by adding an extra parameter in the training dataset. However, this means that extra effort is necessary and again the rating highly depends on who labelled the training dataset. Additionally, other data such as location, the type of rotor blade and age are important.



Overlook of a complete data collection in a 3D environment for easy orientation (one arrow = one data point)

'We continuously improve our knowledge and understanding of what is important every day.'



Anomaly Detection - Using different sensitivity settings

The economic factor

In addition to severity class, it is necessary to consider the economic factor.

For instance: what would be the consequences if certain damages were not identified? Or, is a repair economical: will the operator pay for it?

At first glance it may look like there is no difference between paying attention to these cost functions and the severity grade. Following a five-grade severity class system, it is obvious that damages of categories four and five would have a huge impact if they are missed during inspection. Even if repairing heavy damage on old turbines might not be economical, it is essential that an AI is capable of identifying it.

Being aware that these two categories just contain about 2% of all findings, what is the case with lower severity anomalies? Customers are often not willing to send out a repair team just to fix 'small' damages. Hence, in terms of efficiency, the system should be aware of different cost functions.

At the end there is no single system with one single output that fits all needs. A set of tools, access to intermediate results and the use of different filters are required to cater for the respective needs. It is like in an autonomous car: just recognizing traffic signs is not enough. There are many subsystems which need to be unified in the right way.

Conclusion

Aero Enterprise created an AI-concept based on a huge database, different neuronal networks and logical operations, which is maintained consistently and behaves like a living organism. We continuously improve our knowledge and understanding of what is important every day. Training datasets have been reworked more than ten times to fit the experiences of experts with different backgrounds and also to fulfil customers' needs.

To make reliable statements, human interaction is still needed in the end. With the support of AI, however, we are better able to evaluate the data efficiently. With the increasingly larger inspection projects, evaluation can thus be carried out much faster and, above all, in a standardized manner. The next goal is to further reduce the 'human factor', that is, to further digitize and automate, but not to eliminate the skilled person´s opinion. The final decision is made by the expert.

www.aero-enterprise.com