QUALITY
MATCH

WHITE PAPER:

CAN LAYPEOPLE ANNOTATE MEDICAL DATA?

↘ AN EXPERIMENTAL PROJECT BY CARL ZEISS MEDITEC AG & QUALITY MATCH

For predeveloping prototypes in the field of microsurgery, Carl Zeiss Meditec AG also uses machine learning methods. Here, it is often not the development of code and training model that proves to be the greatest challenge, but rather the selection and annotation of the training data. In this case, the medical technology experts were faced with an immense number of image files to be evaluated as part of a new project, which would have taken them over 200 hours of work to annotate on their own. So it was obvious to get the help of a service provider.

Quality Match is profoundly experienced in the field of quality control of datasets for machine learning, especially — but not only — in the field of computer vision. Quality Match can trump with the “time” factor in particular. Through controlled crowd-based data annotation, labor time can be reduced to a fraction.

But can this also work in the medical field? Can laypeople annotate medical data?

The easiest way to find out is to follow the principle: just do it! This is how Carl Zeiss Meditec AG and Quality Match got involved in this experiment.

ABOUT QUALITY MATCH

Since 2019, the startup Quality Match has been helping companies improve dataset quality through highly specialized decision trees — an annotation method that optimizes the dataset for model training. The objective is to make the dataset as representative, accurate and difficult as possible by filtering out errors, inconsistencies and ambiguities. It is the company’s unique selling point that it breaks down complex quality control tasks into a series of trivial tests that can be performed by untrained annotators.

↘ ADVANTAGES OF CROWD ANNOTATION

Manual annotation took about five minutes per image for Carl Zeiss Meditec AG. Such data sets can be annotated much faster in the crowd-based approach of Quality Match. The annotators do not need to have a medical degree, as long as the annotation tasks are broken down into very simple questions. An example is the question whether human tissue is visible in the image. This can be answered with yes or no and then used in combination with other questions to better understand the context of the image. In this method, the process is continuously improved by quickly capturing errors and anomalies in order to learn from them. Namely, if different annotators give different answers to the same images, these images need to be reviewed by experts. Examining these ambivalent cases can provide information on whether the question should be adjusted, or the image material should be sorted out.

↘ EXPERIMENTAL SETUP

Surgical instruments had to be annotated in separate frames from surgical videos. In order to avoid a time-consuming pre-selection of the images (“Are there any instruments in the image?”, “Does the image even show an OP scene?”), a two-step annotation process was chosen using two so-called annotation pipelines:

- In the first pipeline, the crowd was asked for image attributes, such as the presence of instruments, image sharpness, etc., which were then to serve as filtering and sorting criteria for the second pipeline.
- The second pipeline then dealt with the actual annotation of the tips of the surgical instruments.

In any case, Quality Match takes a very thorough approach to generating annotation pipelines: first, hypotheses are made as to which questions and instructions are particularly effective in achieving the objective. This is checked with the company’s experts and then tested on a small sample data set. Manual sighting and iterative adjustment of the pipeline thus yields better and better results over time. Finally, the pipeline is verified with experts to ensure that it meets the customer’s quality requirements: this is where the optimal trade-off in terms of data quantity and quality can be made on quantitatively validated data.

↘ WHAT ARE THE ADVANTAGES OF SUCH A PROCEDURE?

This method has numerous advantages resulting from the fact that a complex task is broken down into many small tasks:

- Incremental improvements are made easier because only small parts of the annotation pipeline are iteratively adjusted to fit the data.
- Annotation instructions become easier, and laypeople only have to make simple yes/no decisions in many cases.
- Costs can be saved because not every annotation necessarily needs to be performed: depending on what information has already been gathered, one may decide, for example, to simply not annotate the corresponding image at all.
- Small-part attributes can be used for later performance analysis of the AI model (e.g., did the algorithm perform particularly well/badly on blurry images).
- Attributes can be gradually expanded with additional attributes to gather more information about the dataset without having to annotate completely from scratch.

↘ PIPELINE 1:

Data curation via simple questions for each image

The first version of the questions to the crowd was developed in close collaboration, combining the expert knowledge on both sides. The greatest challenge was to convey the expert medical knowledge in such a way that the annotation tasks were also unambiguous for laypersons. In the first pipeline, the initial task was to find a suitable selection of frames (curation). For this, the annotators were asked, for example, about the exposure, the sharpness and the visibility of tool tips.

A good example is the question about the degree of sharpness, which proved to be a particular challenge. Here, with the aid of a continuous scale and a sample image, it was first clarified when an image is considered sharp or blurred. Using a slider and a reference image, the degree of sharpness of the frame was to be classified on a scale of 1 to 10. There was also the option to click on “everything is sharp”, which resulted in a sharpness level of 11.

↘ PIPELINE 2:

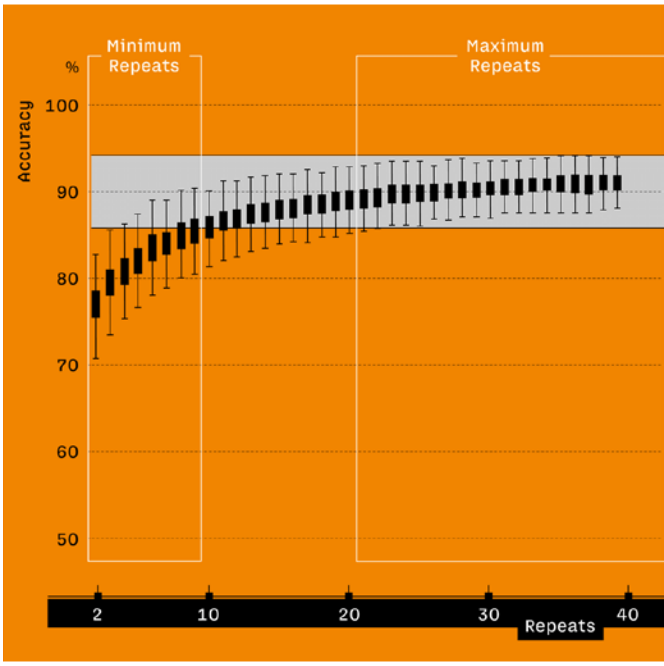
Marking the tool tips in the selected images

In the second pipeline, a polygon was to be drawn around the tips of the instruments, using the same color for the ends of the same instrument, but different colors for different tool tips. Drawing a polygon around an object proved to be a precise middle ground between a circle and a square for annotating objects. A particularly important feature was providing context to the frames that were to be annotated: the annotators were able to better place the particular frame in time with the help of a short video excerpt that was provided.

↘ RESULTS

Depending on the attribute to be evaluated, the results of this first annotation ranged from almost zero errors, as in the case of the question about visible body tissue as a clear objective criterion, to 12 % in the case of the classification of image sharpness, a subjective criterion. For the latter, it turned out that with the given instructions it is apparently difficult to annotate blur with a simple scale: only 74 % of the answers could withstand manual control, 9 % were classified as borderline cases and 12 % as errors. These values are unacceptable for successful model training. Comparison of the annotators with each other showed that they all had almost identical error proportions and deviation values. This indicates that the task can nevertheless be performed with laypersons.

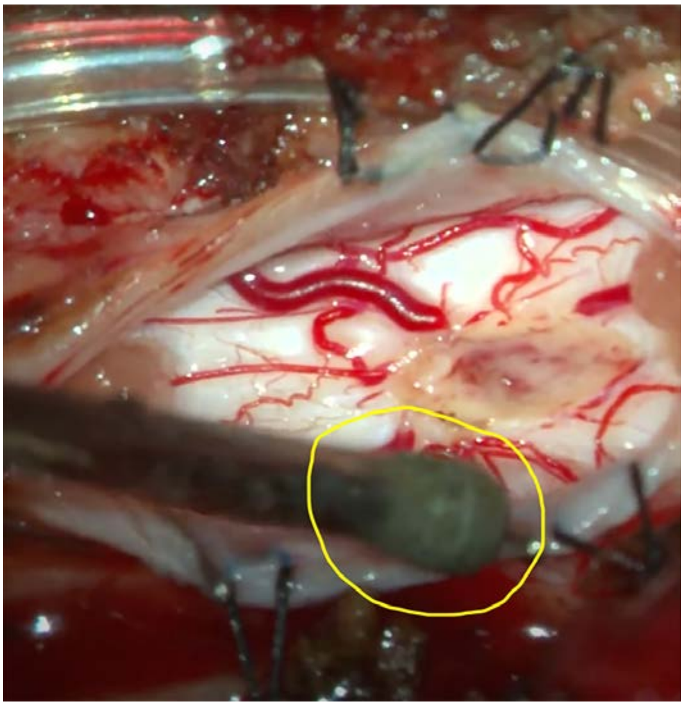
So how can the question be improved to reduce the error rate? Closer examination revealed that the difficulty of the task was that, due to the wide-open camera aperture, different parts of the image were either sharp or blurred. However, the annotators also indicated a lower degree of sharpness if only the edges of the image were blurred, the image would still have been useful for the model. The solution to this problem will be to highlight the relevant area of the image and clearly state in the question that the degree of sharpness should only be given for that area. This is an example of how Quality Match iteratively refines quality control processes for datasets and gradually optimizes datasets using both qualitative, visual hypothesis generation and quantitative, statistical methods.



Annotation accuracy in relationship to number of repeats. Source: Quality Match

Drawing polygons around the ends of various surgical instruments was performed correctly 79% of the time, with 15% of the annotated frames classified as ambiguous cases and 6% as errors. In evaluating these borderline cases and errors, the degree of sharpness stood out as the biggest obstacle to detecting and accurately marking tool tips. It was difficult to detect the ends when the image was uniformly blurred. However, when the image contained both sharp and blurred areas, the degree of sharpness of the tool tips was a key factor in the detection and accuracy of the polygons.

Frame exposure also showed an effect on the ambiguity of responses, but much weaker than for image sharpness.



Polygon around a tooltip, source of the data: https://www.youtube.com/watch?v=KYvcD7Sh-zU&list=PLYDyV8hKe-P3iov3pB-MeaYfOXrdRdGOlm&index=31

↘ OUTLOOK

The collaboration experiment was a great success. Already in the first run, a clear majority of images were correctly annotated with homogeneous annotator ability. With the help of the well-posed and constantly evaluated questions, the cognitive load of the annotators could be reduced, making the annotation of complex data possible for medical laypersons.

As an experienced service provider for quality control of datasets, Quality Match recognized the particular importance of blurriness and processed it for crowd-based annotation, thus making the dataset representative. The inclusion of a small video sequence for the temporal context of the frame also proved helpful and, in some cases, narrowed down ambiguity. This feature will also be helpful in future annotations.

In order to make the conclusions from this project usable for other annotation tasks and use cases, Quality Match is developing the HARI (highly actionable real-time insights) software, in which large visual datasets can be examined and corrected with respect to their quality attributes. A direct link to the annotators enables a smooth, iterative improvement of the dataset. In addition, results of the AI method can subsequently be uploaded to HARI for comparison in order to identify weaknesses in the model and correct them by further improving the dataset.

↘

WE CAN VALIDATE ANY VISUALIZABLE DATASET WITH THIS METHOD TO SEE IF IT IS REPRESENTATIVE, ACCURATE AND DIFFICULT.

IF YOU WANT TO DETECT BORDERLINE CASES AND ERRORS IN YOUR DATASET, CONTACT US TO SCHEDULE A DEMO.

↘

AN EXPERIMENTAL PROJECT BY
CARL ZEISS MEDITEC AG & QUALITY MATCH

Quality Match GmbH
Goldschmidtstraße 5
69115 Heidelberg
Germany

Email: info@quality-match.com
quality-match.com

QUALITY
MATCH