Description of the blank image detection component

Background, benefits and purpose of use of the component

Double-sided scanning is often used in the process of digitising documents. If the documents are one-sided, many blank images are created during digitisation. Machine classification into blank and content images facilitates the digitiser's work, and if the quality of identification is good enough, it also makes it possible to hide or even remove blank images from storage. If the documents are viewable by the customer in the service, the possibility to hide blank images from the view can help improve the customer experience. If blank images are identified before the subsequent stages of document processing (e.g. error detection, optical character recognition, metadata detection), the whole document processing chain can also be accelerated.

What does the component do?

The component processes the image files received as input one at a time and forms a prediction of the content of the image based on the parameters it has learned during the training. If the component identifies content in the image with a probability of more than 50%, the image is classified as having content, and otherwise blank. The component returns the predicted class to the user.

How is the component used?

The component is available both as part of the Arkkiivi user interface (http://www.arkkiivi.fi/) and as a standalone application offering an application programming interface (API) on the DALAI project's GitHub page (https://github.com/DALAI-project/EmptyAPI). More detailed instructions on how to use the component can be found on the attached website.

Component training

Image files in .jpg format aligned to 224 x 224 pixels have been used for component training. The component has been trained several times, and the exact amount of original training material is not known. The material used in the additional training includes approximately 100,000 blank images and 130,000 images with content.