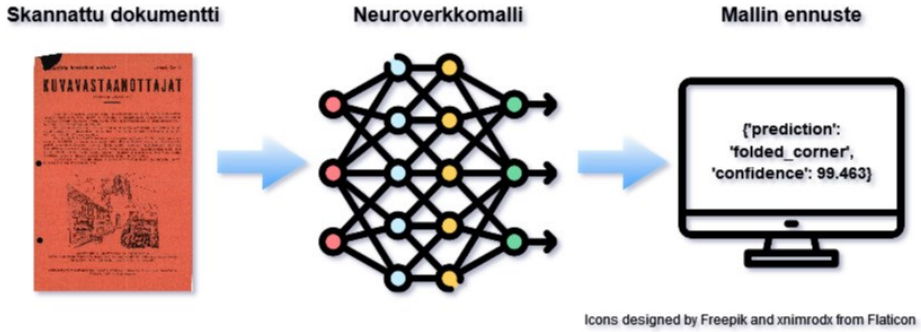
Taittuneiden kulmien tunnistus -komponentti

Komponentin tausta, hyöty ja käyttötarkoitus

Asiakirjojen digitointiprosessissa skannatuksi saattaa tulla asiakirjoja, joissa paperin kulma tai reuna on taittunut tai revennyt. Mikäli taitos tai repeämä vaikuttaa asiakirjan sisällön luettavuuteen, on hyvä, jos tällaiset tapaukset voidaan tunnistaa ja tarvittaessa skannata uudelleen. Taittuneiden kulmien koneellinen tunnistus auttaa näin parantamaan digitaalisen aineiston laatua ja vähentää tarvetta manuaaliseen laaduntarkistukseen.

Mitä komponentti tekee?



Komponentti käsittelee syötteenä saamansa kuvatiedostot yksi kerrallaan ja muodostaa koulutuksen aikana oppimiensa parametrien pohjalta ennustuksen kuvan sisällöstä. Mikäli komponentti tunnistaa kuvasta taitoksen tai repeämän yli 50 % todennäköisyydellä, kuva luokitellaan virheelliseksi. Komponentti palauttaa käyttäjälle ennustetun luokan.

Miten komponenttia käytetään?

Komponentti on saatavilla sekä osana Arkkiivi-käyttöliittymää (http://www.arkkiivi.fi/) että itsenäisenä, ohjelmointirajapinnan (API) tarjoavana sovelluksena DALAI-hankkeen GitHub-sivulla (https://github.com/DALAI-project/CornerAPI). Tarkemmat ohjeet komponentin käyttöön eri ympäristöissä löytyvät oheisilta verkkosivuilta.

Komponentin koulutus

Komponentin koulutukseen on käytetty .jpg-muotoisia kuvatiedostoja, joiden koko on yhdenmukaistettu 224 x 224 pikseliin. Koulutuksessa käytettyjä dokumentteja on yhteensä n. 35 000 kappaletta, joista taitoksia tai repeämiä sisältäviä dokumentteja on n. 5 000. Koulutusaineisto on valikoitu niin, että komponentti tunnistaisi dokumenteista taittuneita ja revenneitä kulmia riippumatta siitä, vaikuttavatko virheet suoraan dokumentin sisällön luettavuuteen (esim. taittuneen kulman ei tarvitse peittää tekstisisältöä).

Vaikka koulutusaineistoon on pyritty kokoamaan monipuolisesti erilaisia esimerkkitapauksia, aineiston määrä on rajallinen ja komponentin luokittelussa tapahtuu myös virheitä. Esimerkiksi mikäli dokumentin kulma muistuttaa värityksensä tai muotonsa takia taitosta, saattaa tämä johtaa virheelliseen luokitukseen.

Folded Corner Detection Component

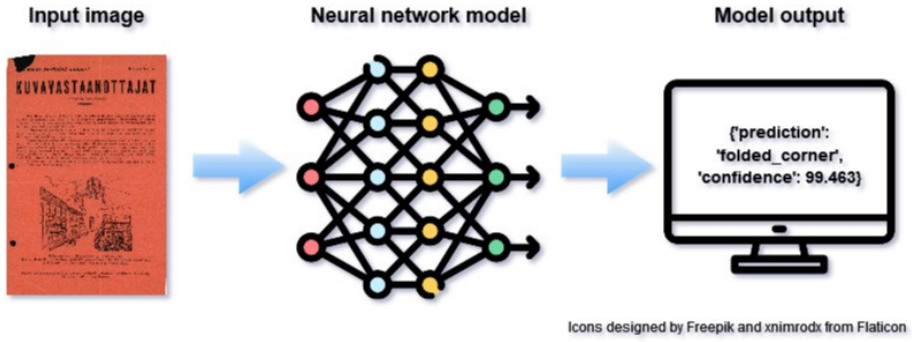
Background, benefits and purpose of use of the component

Documents in which the corner or edge of the paper is folded or torn may be scanned during the document digitisation process. If the fold or tear affects the readability of the content of the document, it is useful if such cases can be identified and, if necessary, scanned again. Machine identification of folded corners thus helps improve the quality of digital material and reduces the need for manual quality control.

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 a fold or tear in the image with a probability of more than 50%, the image is classified as incorrect. 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/CornerAPI). More detailed instructions on how to use the component in different environments 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. A total of about 35,000 documents have been used in the training, of which approximately 5,000 are documents containing folds or tears. The training material has been selected so that the component would identify folded and torn corners of the documents, regardless of whether the errors directly affect the readability of the document's content (e.g. a folded corner need not cover the text content).

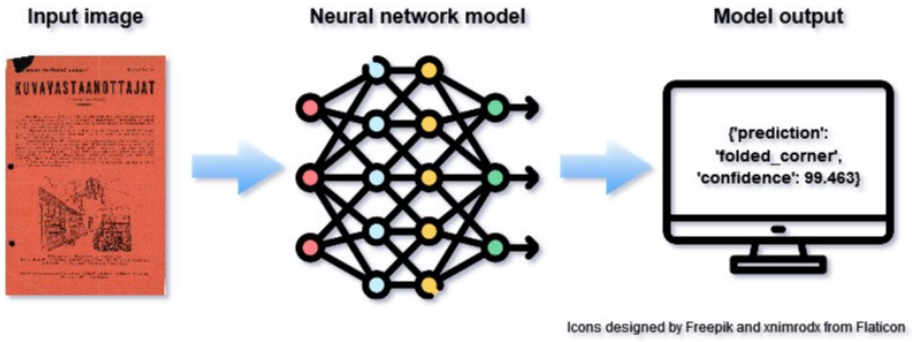
Although efforts have been made to compile a variety of example cases for the training material, the amount of the material is limited and errors also occur in component classification. Among others, a document corner resembling a fold due to its colour or shape may lead to an incorrect classification.

Igenkänning av vikta sidhörn

Bakgrund, fördelar och ändamål med komponenten

I processen för digitalisering av dokument kan det hända att dokument där ett hörn eller en kant på pappret är vikt eller sönderrivet överlämnas för skanning. Om det vikta eller sönderrivna sidhörnet påverkar läsbarheten av innehållet, är det bra om sådana sidor kan kännas igen och vid behov skannas på nytt. En maskinell igenkänning av vikta sidhörn bidrar därmed till att förbättra kvaliteten på digitalt material och minskar behovet av en manuell kvalitetskontroll.

Vad gör komponenten?



Komponenten behandlar de bildfiler som den fått som indata en efter en och tar fram en prognos för bildens innehåll på grundval av de parametrar som den har lärt sig under utbildningen. Om komponenten känner igen en vikt eller sönderriven bild med mer än 50 procents sannolikhet, klassificeras bilden som felaktig. Komponenten återställer den förväntade klassen till användaren.

Hur används komponenten?

Komponenten finns tillgänglig både som en del av Arkkiivi-gränssnittet (http://www.arkkiivi.fi/) och som en fristående app som tillhandahåller ett programmeringsgränssnitt (API) på DALAI-projektets GitHub-webbplats (https://github.com/DALAI-project/CornerAPI). Närmare anvisningar om användningen av komponenten inom olika miljöer finns på den webbplatsen.

Utbildning av komponenten

Bildfiler i .jpg-format har använts för att utbilda komponenten och deras storlek har harmoniserats till 224 x 224 pixlar. Det totala antalet dokument som använts i utbildningen uppgår till cirka 35 000 exemplar, av vilka cirka 5 000 är vikta eller sönderrivna dokument. Utbildningsmaterialet är utvalt så att komponenten kan känna igen vikta eller sönderrivna sidhörn i dokumenten, oberoende av om felen påverkar läsbarheten av dokumentet direkt (t.ex. behöver ett vikt sidhörn inte dölja textinnehållet).

Trots att man har försökt samla många olika exempel till utbildningsmaterialet, är materialmängden begränsad och det uppstår också fel i komponentens klassificering. Om ett dokumenthörn till exempel på grund av sin färg eller form liknar ett vikt sidhörn, kan detta leda till en felaktig klassificering.