Team W14: Hain Zuppur, Alvin Meltsov

Task 2: Business understanding

Identifying your business goals:

Our project is done in cooperation with the historical pictures repository Ajapaik, which in addition to accessing data from various national archive databases also expands on the available data by offering user-friendly website. On this website anyone can add pictures, mark faces, check duplicate photos or browse databases through time in spatial maps provided by Google. This is especially useful for either researchers for seeing data from different databases in one focused site or for occasional users who do not know all the available sources to check for. As a database, duplicate pictures do not serve any purpose except for quality improvement or for mapping metadata. Therefore, our goal is to diminish the extra space wasted on duplicated photos and by helping with the detection and extraction of unique pictures from different photos or scans of old photos. In conclusion, our project has to provide an algorithm / data mining method, that can successfully detect the frames or extra space from different excerpts of the pictures for the downstream inhouse picture duplicate detection, during December 2019.

Assessing your situation:

To fulfil our goals we have:

1) Data from the Ajapaik's collection of archived historical pictures, from which we have selected some prime examples of framed captions of various unpurity. For those pictures we already have some similar pictures to compare them with, but our goal is to only detect and mark the frame and crop the picture accordingly.
2) Data from team members private collection of old pictures, which are also added to Ajapaik's repository during this project and can also serve as testing data to our algorithm.
3) Ajapaik's internal duplication detection algorithm that is successfully working already and is available to use from Github downstream for our algorithm (construction of which is not our project's main goal).
4) Willing team members with capable computers and an access to the internet.

Previous works: we have searched but have not found any relatable prior work for exactly this same problem, therefore we must work out our novel solution on our own.

Risks: Our main concern with this project is that we can not find an effective way to detect edges of frames from our pictures. If we are able to detect frames in some good cases, then further problems might arise from:

- **doubled-pictures** where two pictures are in one frame;
- **unaligned pictures** where the frame of the real picture is not perpedicular to the edges of the scan (edge of digital picture);
- **low quality pictures** where the quality of the picture distrupts from successfully detecting the edge and results in erroneous croppings;
- **photos of photos** which are photographs or scans of a photo album in real life or another photograph in bad lighting conditions.

Defining your data-mining goals:

Our main goal is to create a working image-in-image detector for defining the frames from best-case scans of old photographs. As there are different kinds of problems with historical imagery, then our subsequent goals are based on those challenges (doubled-pictures, unaligned pictures, low quality pictures), which are specified in the **Risks** section.

Task 3: Data understanding

Data requirements:

Our data is archive pictures and the images must have a photo border. We don't have strict requirements for our image quality, because we will try to get our algorithm/model to work on wide range of different resolutions. Our images are in “png” format.

Data availability:

We get our data from ajapaik.ee and from Alvin Meltsov personal archive of Navi village. In Ajapaik many pictures have a border on them and they don't have any tools to automatically crop them, so from there we can get lots of pictures for testing our algorithm/model. If we would need some more data then Hain Zuppur has also a family archive of pictures with frames that we could use. There are lots of suitable images to test our algorithm/model on in different archives, and most of them are freely accessible.

Selection criteria:

We select pictures with clear and distinguishable photo border. We will try to find pictures with many different styled borders, some have 90° edges and some have round edges. For our algorithm/model the content of the picture is not important.

Describing data:

Our data is archive pictures with different photo border. For our initial tests we downloaded 21 pictures from Ajapaik and manually found the coordinates where correct place would be for cropping. The pictures are usually portraits or pictures of buildings. Most of the images are black and white, but some are colored.

Exploring data:
From our initial data set of 21 pictures we found some potential issues, some of the images have low contrast and it is hard to distingues where the picture ends and the border starts. Another potential issue is that some pictures have photographed at an angel and so the image is slightly warped. The images have also wildly varying borders, some are plain and some are decorated. Also some images have text written on its border, some have rulers on the side, some have color card in the side and some have archive watermark on them. The images resolution is mostly around 800 x 1200 pixels.

Verifying data quality:
For out initial data set we tried to draw the cropping rectangle on the image manually to see if the photo is straight and is not warped. From that we discovered that in some of the images the some of the photo is slightly out of frame and because of this the image might be unusable.

Task 4: Planning

1) Contact Ajapaik and get an overview of possible solutions and problems. Alvin and Hain; 3h each (done)
2) Set up Github project page and collect good test pictures on which to develop and test our algorithm. Hain; 3h (done)
3) Manually find the coordinates for the picture frames in the initial data set. Hain 3h (done)
4) Search for possible known algorithms or methods regarding our data and our goal. Alvin 3h; (done)
5) Develop the algorithm for detecting frames in best-case scenarios. Alvin, Hain; 15h each.
6) Explore the different bad-case scenarios of picture frame detection (optional, depends on previous step). Alvin; 9h.
7) Experiment with the pipeline by cropping the pictures with the algorithm and searching for duplicates using Ajapaik’s website. Hain; 6h.