Assignment 1

Team members and their roles:

Customer: Sirojiddin Komolov

Dataset Team: Demyan Zverev, Munir Khisamov

Detector Team: Anastasia Pichugina, Nikita Rashkin

Classifier Team: Emil Davlityarov

Team members: Emil Davlityarov, Demyan Zverev, Nikita Rashkin, Anastasia Pichugina, Munir Khisamov.

Interview Script:

A few days before the meeting, we met with the team and discussed the project and its implementation. We began to think about how we would implement this project. We discussed ideas, possible role distribution and tried to find similar projects that have a similar idea. After analyzing and deriving the criteria for evaluating each of them: the accuracy of determining the seal, the volume of the database on which the machine was trained, the quality of the classification, work with a variety of seals. We also thought about the platform where this project will be implemented and brought this thought as a question. We searched for databases of documents containing seals and made sure that they were indeed in the public domain. Each of us gave a number of questions that interested them, after which they identified a number of the most important of them:

1) What platform will we use for this project?
2) Is there a stack preference? (which libraries, algorithms)
3) What is the input/output format for the documents?
4) Which types of figures would be represented as stamps?
5) How should we classify stamps?
6) Should we use AI to generate documents or is any graphical library ok?
7) Will there be overlapping seals? How many stamps would consist of a document?

8) Do we need to train models such that it recognizes the particular set of stamps or just classifies their shape?

9) What languages do the documents use?

Research existing solutions:

Characteristic\Link	https://www.hindawi.com/journals/mpe/2015/367879/ (Advanced AI model)	https://vc.ru/newtechaudit/611937-raspoznavanie-kruglyh-pechatey (Very simple Hough transform approach with manually derived coefficients)	https://www.researchgate.net/publication/224265585_Stamp_Detection_in_Color_Document_Images (Model with manually derived formulas and coefficients)
Accuracy of detection	70.44%	81%	83%
Size of the database	400 documents	-	400 documents
Accuracy of classification	Approximately 98% (number are a bit different for some shapes)	-	-
Overlaying stamps detection	The model has some problems, but it supports such functionality	-	Model has an acceptable accuracy of 69% in such complex scenarios

All the existing approaches use the YCbCr color model instead of RGB.

1) https://www.hindawi.com/journals/mpe/2015/367879/
To detect circular stamps CHT(Circular Hough Transforms) are used. Image is binarised, Canny filtering is performed, and noise is removed before using CHT. A separate complex approach is used to detect figures, consisting of straight lines. Several parameters are evaluated in order to get the shape of the detected stamp, like roundness, squareness and others.
2) https://vc.ru/newtechaudit/611937-raspoznavanie-kruglyh-pechatey
To detect stamps on the list, we should determine the figures on the list. This project identifies the presence of round seals on the page of the scanned document. It uses Python language and several useful libraries, such as: open source, matplotlib.pyplot, numpy, imutils , which will help us in our project. Also Hough Transformation used to parametrically identify the geometric elements of a raster image. The disadvantages of this project are: absence of database, classification of stamps. But a good feature is high accuracy detection is 81%.
3) https://www.researchgate.net/publication/224265585_Stamp_Detection_in_Color_Document_Images

This model is based on color ink cloud elongated shapes detection. It works well with logos, as authors detect them looking at nearby color areas. Both 200 and 300 ppi documents give high accuracy.

Interview notes/transcript;

We made an appointment for 3pm on Sunday. It went online on Zoom. Our entire team and customers were present. By agreement with the customer, we communicated in Russian. At the beginning, we talked about ourselves, about our main skills and about our experience with computer vision technology. After that, the customer announced that due to the fact that many teams chose his project, he prepared another one, which is very similar to the original, but still with core features. It was a signature recognition project. Due to the similarity of the projects, we decided that we would make a final decision after the meeting, and the list of questions did not change much, so we were able to ask all the questions we were interested in and received detailed answers to them:

1) What platform will we use for this project? Do we need to make some kind of front-end for this project?
Answer: no need

2) Is there a stack preference? (which libraries, algorithms)
Answer: There is no specific stack, but it is recommended to use python

3) What is the input/output format for the documents?
Answer: Image with varying resolution starting from 200 ppi

4) Which types of figures would be represented as stamps?
Answer: various forms, the program must be able to recognize all

5) How should we classify stamps?
Answer: the customer explained to us the actual application of this project, on the basis of which it became clear what is meant by classification. The real application is as follows: some company wants to know about the sending of their documents from the company and for this they need a program that will be able to find stamps on documents and determine whether they are stamps or not. (He also answered the 8th question with this answer)

6) Should we use AI to generate documents or is any graphical library ok?
Answer: it will be enough to have an algorithm that will generate a document with printing in a random place, there is no need to train a model or use AI for these purposes

7) Will there be overlapping seals? How many stamps would consist of a document?
Answer: To begin with, we will not consider the case of intersecting seals, but in the future we may return to this topic, it all depends on how successfully our team will cope with the project. The number of stamps on the document is not limited

9) What languages do the documents use?
Answer: in the implementation of our project, which the customer imagines, there is no need to know in which language to write the document

After asking questions and receiving answers, the customer noted that he believes that this is more of a research project, he wants to see what problems we will face and how we will solve them, how quickly we will find and analyze information.
At the end of the meeting, the team decided that we would like to do the initial project to determine the seals on the documents.

Report:

1) From a conversation with the customer, we determined that this project should be created to recognize classified and genuine documents. This goal can be pursued by companies to prevent data leakage. The main tasks will be the detection and classification of stamps. Moreover, no matter what language the document uses. The model only must distinguish the text and the stamp. Besides, the customer advised us to use python for it. Our task is to train a model, and no design representation is required.

2) After studying the available research, we found which python libraries are better to use for such a project. We also found out that machine learning models for detection do not support seal classification well, which is what we have to do.

3) After discussing the goals and ways of dealing with this project, some questions appeared about generating documents:
    - Should we use stamps appropriate for this document on the subject or the meaning of the seal and the document are not important?
    - How many different stamps should be applied to documents? Will 20 stamps be enough for this?
    - Is it possible to take a ready-made dataset of documents and print over them?
    - How to differentiate “stamps” and “not stamps” for classification?

4) The next step is to split the big task into several smaller ones to separate the responsibilities. Everyone must choose in which area he will work. At first glance, there are 3 main topics: dataset generation, a model for detection, and the implementation of stamp classification.
   For the most efficient execution of a project, we should firstly focus on creation of a dataset for analysis. After that, we can move to detection and classification problems. The model must determine whether there is an object similar to a stamp on the document, how many such objects are. We also initially assume that no intersections of seals are implied. Next, it needs to find out their location, for each to display the coordinates