Structured Text Extraction from ID Documents

Olti Qirici[1]

[1] *Faculty of Natural Sciences, University of Tirana, Tirana, Albania*

Abstract

It will be described hereafter an experiment comprised of the methodology and results obtained for which not many years ago, entire PhDs were written and the results unsuccessfully described. The Generative AI revolution made it easy that through small amounts of code, with limited resources, can achieve nearly instantaneously a result that years ago was considered state-of-the-art. Companies driving such services have faced a sudden fall in incomes and even are threatening to close business. While this is not an economics analysis but a technological assertion of our time, let us utilize these technological advancements to illustrate the easiness in which the extraction of text from image, even on complex images such as ID Cards, which have watermarks and color differences, can be performed in a simple and more affordable manner. In the following paper, the author will be describing the results of an experiment on using Generative AI services to extract from specimens of ID cards from various European countries, the details of the ID Card, and furthermore, test the same on smart phone taken photos, under low quality lighting and receive the same results. The text information will be extracted in JSON format ready to be published by any web service.

Keywords

GenAI, Gemini Flash 2.0, AI Agent, Text extraction, ID Cards information

1. Introduction

For centuries humans have tried to vest into machines, human capabilities. This has been the dream of millennia. Speaking, listening, seeing, sensing machines that do hard work for humans. While these aims to have machines instead of men that do the hard labors have accompanied the humanity for ages, it would be the privilege of the 20th and most notably the 21st century that would literally do a stupendous leap into this direction.

We couldn't but notice the tremendous advancement brought by a specific such achievement, which enlightened the post-covid period and indeed made it possible for this state-of-the-art technology to disrupt the way computers were behaving into a new field of communication. This breakthrough is the GenAI. There have been for some decades attempts to build such machines, but it would be the privilege of the 20th decade of the 21st century to build such a marvelous technology.

During a discussion in our Computer Science Department, on the achievement of GenAI, the author emphasized that years ago, doctorates with dubious results would have been written on the matter we are going to present, the text extraction from an image. Similar solutions such as extraction from PDF documents, as distilled PDFs inclusive of images have been reported in the literature [1]. And while there were methods from filtering an image upon finding edges on it (please refer to a fairly close in time paper written by Kumar R. et al. [2]), to utilizing the transformed image as an input to a pretrained pattern recognition system, such as a Neural Network, the PhD candidate would try to achieve the desired results sometime by meticulously reproducing the very same font type as the one used to train the system. [3] Such a challenge was described by different authors such as also by Sumathi et al [4] which despite the different methods used still notify the readers of the difficulties

extracting the text from the image. Nowadays, this is not needed anymore. Utilizing a simple GenAI service and calling the required information with an adequate prompt would produce far better results in a shorter time.

We will try to build hereafter a system that is based on two basic GenAI calls towards the Gemini API interface and extract the text from specimen documents of various ID cards and then produce a JSON which might then feed an API Endpoint to organize this very data. This method might achieve fabulous results especially when the extraction of such data would be supervised by a human which corrects, and updates wrongly extracted values.

2. Methodology

The methodology used would utilize the service of Generative AI to extract information on the ID Cards of various European countries. In order to have a broader example we will try languages which not necessarily have Latin letters such as the Greek ID Card, specific languages of the European Branch such as the Albanian Language, and other ID Cards such as the Kosovo ID Card and the Croatian ID Card. For all these ID Cards the specimens will be utilized to extract the information (the specimens of ID Cards are general ID Card patterns for a specific country that define the visual standard of such a document).

After choosing the input information a model should be specified. The model that was chosen for GenAI is Gemini Flash 2.0. As identified by Gemini [5], the Gemini API offers different models that are optimized for specific use cases. And more specifically Gemini Flash 2.0 is optimized for Next generation features, speed, thinking, real-time streaming, and multimodal generation. [5]

An additional difficulty that comes from the same source it is that the Gemini (at least on the day of the experimentation), does not support the Albanian language. But as we'll see this is not a problem at least for the kind of problem that we'll try to solve.

According to Charles Braque [6], whilst text extraction from pdfs can leverage natural language processing and large language model technologies, text extraction from images relies more heavily on computer vision and optical character recognition. Anyway, we will show that the results from models such as Gemini Flash 2.0 are indeed impressive. This is something that has been anticipated in various technological environments. As Regalado D. expressed himself: "This is a huge leap forward, not just in model performance but in AI's ability to be more integrated and effective, it is a big step towards more powerful AI tools for all of us." [7] Another advantage of using Gemini Flash 2.0 is that Google offers limited utilization of APIs for Enterprise Google Users, which makes it easy and cost efficiently for experimentation and for didactical purposes.

 In the following image a system has been designed for this specific experiment:
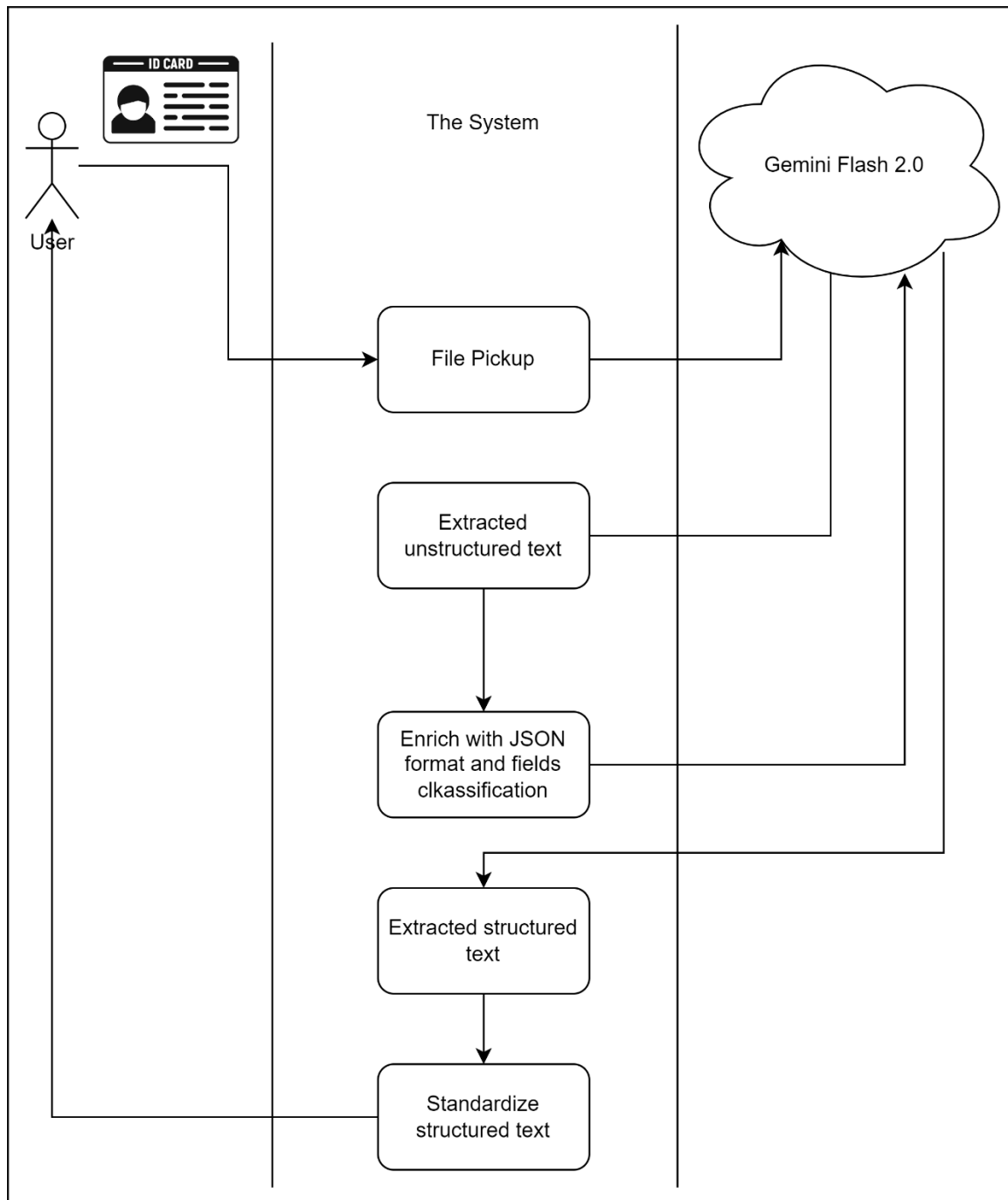
Figure 1: Extraction of information through AI from ID Cards

As can be seen from the schema, the ID Card that the user will upload will be posted into the Gemini Flash 2.0 API with the instruction for text extraction. Following the receipt of the text from the cloud service, the text should be formatted in the specified JSON format so another call is sent toward the API which would try to allocate the information to the respective field. For this purpose, the AI service should semantically analyze the text and allocate the appropriate information to the appropriate JSON label.

In order to have a simpler way to delve into the design of the system than to program the parts that are general, Python was chosen as the programming language. Simple snippets to link with Gemini API services are available across the web.

Workshop organizers may want to provide a copy of this template to authors where the event title in the footnote is updated to their workshop details, see "Woodstock …" footnote on page 1. While we provide a Word/LibreOffice template, we strongly recommend authors to use our LaTeX template.

3. Results

As previously explained, the first step of the experiment is the acquiring of ID Cards for some European countries. In the following image there are shown some samples (specimens) of real ID Cards utilized for experimentations (all specimens were extracted from the specific Wikipedia pages of the ID Cards of each of the above countries):

Figure 2: Specimens for the Albania, Kosovo, Greece, Croatia ID Cards (tiles from the upper left side to the lower right side).

The above cards were used for experimentation.

The extract from the system after executing the first step, the step related to the text extraction (and hereafter I'm demonstrating the solution only for the Albanian ID Card, by emphasizing that the same results were received also by executing the code for the ID Cards of other countries) is the following:

```
LETËRNJOFTIM ID
REPUBLIKA E SHQIPËRISË REPUBLIC OF ALBANIA
mbiemri/surname: Sojli
emri/given name: Monika
shtetësia/nationality: Shqiptare/Albanian
vendlindja/place of birth: Tirana
datëlindia/date of birth: 01-01-1980
data e lëshimit/date of issue: 16-06-2004
autoriteti lëshues/authority: MB
SPECIMEN
nr. letërnjoftim/card no.: 200000907
gjinia/sex: F
data e skadimit/date of expiry: 15-06-2014
nr. personal/personal no.: I05101999Q
ALB
firma/signature: Monika Sojli
```

```json
{
  "card_type": "LETËRNJOFTIM ID",
  "country": "REPUBLIKA E SHQIPËRISË",
  "surname": "Sojli",
  "given_name": "Monika",
  "card_number": "200000907",
  "sex": "F",
  "nationality": "Shqiptare/Albanian",
  "place_of_birth": "Tirana",
  "date_of_birth": "01-01-1980",
  "date_of_issue": "16-06-2004",
  "issuing_authority": "MB",
  "date_of_expiry": "15-06-2014",
  "personal_number": "I05101999Q",
  "signature": "Monika Sojli"
}
```

Figure 3: Extracted unstructured text from the Albanian ID Card (the specimen) on the right, structured information on the left.

Following the above information extraction, a JSON structured template was sent to the Gemini Flash 2.0 API. The structured information was correctly detected and provided. Indeed, if the images above can be thoroughly investigated, it can be identified that all the extracted data were correct (even the handwritten information, such as the signature field).

The results have been checked also on other ID Cards (which cannot be shown for ethical reasons and because of the personal information holding), and the extracted information was always correct.

4. Conclusion

We demonstrated above the ease to implement a solution such as the above proposed to solve the issue of Document Information Extracted. The same experiment was driven for different types of documents and what was achieved was an excellent result for the document extraction (for specific documents such as Water Supply bills or Electricity Bills). Even when the medium was badly handled (the paper washed out in a washing machine or shredded) and the data was hardly or doubtfully read by the human eye, still the Gemini Flash service had the capability to extract the text.

Even though the experiment for the above paper was performed by Gemini Flash, mainly for cost cautiousness reasons, the same results were received by executing the above experiment on ChatGPT, GPT 4.o mini etc. Of course, the Agentic AI (the AI Agents in general) is transforming the technology and as such is utilizing the information produced by the GenAI to plan and act accordingly (so even trying to build a logical map as the above designed itself).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools to create or edit the above text.

References

[1] IBM watsonx, "Extracting text from documents," 26 March 2025. [Online]. Available: https://dataplatform.cloud.ibm.com/docs/content/wsj/analyze-data/fm-api-text-extraction.html?context=wx.

[2] e. a. Kumar R., "A Study: Extraction of Text on Image," *International journal of advance reasearch and Innovative ideas in education, Vol-7, Issue 1, 2021,* pp. 904-909, August 2021.

[3] A. Rosales, "Top AI tools for text extraction," 21 June 2024. [Online]. Available: https://medium.com/@andrea.rosales08/top-ai-tools-to-extract-text-from-documents-43c3641124a2.

[4] S. T. P. N. Sumathi C. P., "Techniques and challenges of automatic text extraction in complex images: A survey," *Journal of Theoretical and Applied Information Technology, Vol.35, N.2,* pp. 225-235, January 2012.

[5] "Gemini models," 27 March 2025. [Online]. Available: https://ai.google.dev/gemini-api/docs/models.

[6] C. Brecque, "An Introduction To Text Extraction From Images," 19 July 2024. [Online]. Available: https://textmine.com/post/an-introduction-to-text-extraction-from-images.

[7] D. Regalado, "Google Gemini 2.0 Flash Is Here!," 12 December 2024. [Online]. Available: https://davidregalado255.medium.com/%EF%B8%8F-gemini-2-0-is-here-c593b6fe6de0.