Lausanne 2024 Abstract Submission

Title
Text based research in an archive of wine labels using Optical Character Recognition

I want to submit an abstract for:
Conference Poster Session

Corresponding Author
Stephan Euler

E-Mail Corresponding Author
Stephan.Euler@mnd.thm.de

Affiliation
Technische Hochschule Mittelhessen

Keywords
wine labels archive, text based requests, optical character recognition

Research Question
Can OCR help to search for wine labels in an archive based on text information?

Methods
A small database of manually transcribed labels is used to examine how often the terms are found in the automatically generated texts.

Results
Depending on the language of the labels we found recall rates between 76% and 92%. A text based search has been integrated into the archive.

Abstract
Wine labels are an important source of information. In addition to the legally prescribed information, which varies from country to country, the design and other texts provide important information. Therefore, a repository of wine labels can be a valuable basis for research.

Over time, we have built up an online archive of almost 25,000 labels (www.weinetikettensammler.de). The focus is on wines from Germany, but there are also numerous labels from other countries. For each label, geographical information, the vintage and some tags are currently stored. In order to increase accessibility for research, access to the text information would be important. This would allow to search e.g. for wineries, grape varieties or historical information.

In view of the large number of labels, automatic text recognition would be desirable. Systems for optical character recognition (OCR) have improved considerably over the time. However, the often very specific design of the labels poses major challenges. There are very different fonts and text sizes, sometimes even in different orientations horizontal and vertical. In some cases, it is not just a single label but a combination with a back label or neck loop. While some of the terms are prescribed (origin, grape varieties, etc.), there are also proper names (name and address of the winery) and in some cases more or less free descriptions. Finally, some labels even feature imaginary names for the wines.

In this article, we report on initial investigations into the extent to which search queries are answered correctly.
using automatically extracted texts. To this end, 10 labels each were initially selected from the USA, Germany and Switzerland (German-speaking). These countries have very different requirements and traditions of label design. For each label, all texts were recorded manually and written to a reference file. The texts were divided into meaningful units. Examples include grape varieties (Riesling) or geographical terms (Sonoma County).

This resulted in an average of 14 entries per label for the German wines, while the American wines had 11 and the Swiss 10. Texts were then extracted from the labels using OCR. We used the two systems Tesseract and EasyOCR. In order to improve the OCR quality, the language was given for each label during processing. Finally, we counted how many of the manually determined entries were found in the OCR generated files. This resulted in the following rates:

- Tesseract: DE 59%, US 81%, CH 67%
- EasyOCT: DE 56%, US 72%, CH 74%

The results for individual labels vary strongly with a minimum of 11%, a maximum of 100% and an average value of 68%. To test whether the results of the two programs might complement each other, the combination of the two results was examined. The results

Tesseract + EasyOCT: DE 76%, US 92%, CH 82%

shows a clear improvement compared to the individual rates. It can be noted that the labels from Germany have the largest text content on the one hand, but are the most difficult to recognize on the other. Overall, the values are encouraging, although the database needs to be enlarged for a more precise analysis.

For a practical test, the search was integrated into the web application. The texts for a subset of the labels were extracted using Tesseract. Currently 3431 labels were included. Recognition failed completely for 215 images (6.3%) and no texts were recognized. In some cases, these were artistically designed labels with corresponding font designs. In other cases, the low contrast between text and background could have been a problem.

In order to gain an overview of the occurring terms, the recognized texts were converted to lower case and then divided into individual words based on the whitespaces. A minimum length of 4 characters was specified. In this sense, a word is a sequence of at least 4 characters without spaces. This also includes, for example, year dates or information such as 75cl.

This resulted in a total of 97,610 words with an average of 30 words per label. The frequency analysis revealed a total of 23,919 different words. Of these, however, 18,932 (79.2%) occurred only once. Since no restrictions other than the minimum length have been applied so far, there are certainly many misrecognitions among them. The website contains a list of all words that occur at least 20 times. As expected, given terms such as Riesling, Qualitätswein or Appellation are the most frequent. The top entries are Wine and France.

The initial results are encouraging. A lot of information can already be extracted automatically from wine labels using OCR. This has already made it possible to enter queries that return at least some of the labels from the archive for the given search term. But there is still a lot to do. First, the OCR process should be examined more closely and optimized. The extension to other languages with their own characters should also be considered. The search is currently limited to a single term. More complex options such as combining several terms or excluding others would be useful. Ultimately, a semantic annotation would be helpful to assign the terms to categories, thereby allowing structured queries.

**Consent**

✔ I agree to the privacy policy.

You find the link to our privacy policy at the bottom of the page. In the privacy policy you find a link for exporting and/or erasing your personal data stored in our database.