Extraction and Classification of Historical Stamp Cards using Computer Vision

KENZO MILLEVILLE, Ghent University - imec, IDLab
KRISHNA KUMAR THIRUKOKARANAM CHANDRASEKAR, Ghent University - imec, IDLab
THIBAULT BLYAU, Ghent University - imec, IDLab
AURORA IANNELLO, University of Turin
UMBERTO MICHELUCCI, TOELT LLC, Switzerland
STEVEN VERSTOCKT, Ghent University - imec, IDLab

1 INTRODUCTION

The digitization of historical documents has led to a surge in digital and cultural heritage research, highlighting the need for automated tools to process and enrich these collections. Annotating them with rich metadata will improve their accessibility and open new research opportunities. This work details a computer vision-based annotation pipeline, with limited manual annotation to facilitate studies of the Italian economy at the beginning of the last century. The existing efforts undertaken in quantitative economic history have made it possible to reconstruct consistent historical series describing its fundamental features and trends. Nowadays, many historical series are available for the interwar period in Italy. These detail the size of the workforce [1], the number of hours worked [12], wages [10, 12], labor productivity [1, 6], and structural changes in employment over sector [3, 11], gender [3], and geographic regions [2, 4, 5].

The project focuses on a previously unexplored source, namely Italian social security cards issued in the 1920s and 1930s, to collect and explore new micro-data on the working life of private-sector wage workers in the interwar period. These historical cards are stored in the provincial documentary archives of the National Institute of Statistics (INPS). They contain detailed information about each worker, describing their name, date, birthplace, occupation, weeks worked, wages, and information about their employer. A sample of 70 stamp cards was selected and digitized from the archives of Vercelli, Turin, Rome, and Naples. Each card contains one page with multiple stamps, denoting the worker's payout for the given week. The card layout, types of stamps, and their monetary value (in Italian lira) vary greatly over the years, making manual annotation time-consuming.

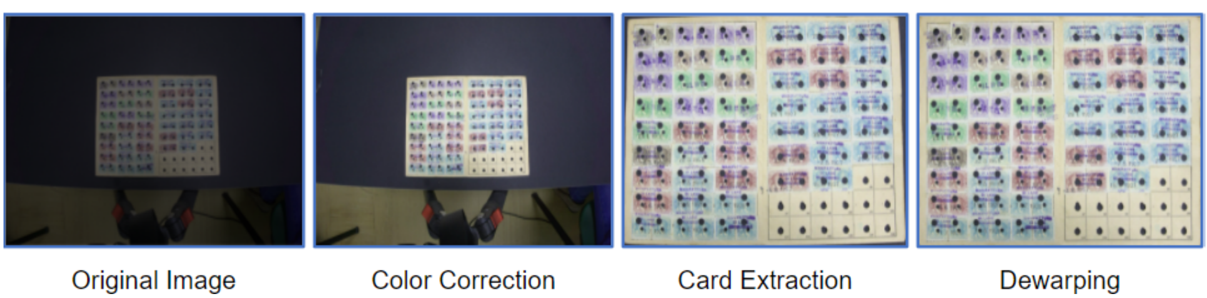

Fig. 1. Overview of the preprocessing pipeline.

2 STAMP DETECTION

First, the digitized cards were preprocessed. After a color correction, the entire cards were segmented from the background via K-means color clustering. Afterward, the cards were dewarped and flattened. The preprocessing pipeline is shown in Figure 1.

Our initial approach to extract the individual stamps from each card was based on a simple grid system. The background color of the cards was used to determine the grid on which the stamps were placed. Extracting the stamps using this grid did not work well, because some pages were not perfectly straight and stamps were often misplaced. Therefore, we moved over to an object detection approach via transfer learning. We used the YOLOv5 model, an improved version of the YOLO architecture [8]. 33 pages were manually labeled with the bounding box of each stamp, totaling 2411 stamps. Three of those pages were used as validation data. The model predictions were then subsequently post-processed. Predictions with a confidence score lower than 0.5 were removed. Predicted stamps that were much larger or smaller than the others were also detected as outliers and removed. Figure 2 shows the output of the YOLO model and post-processing results. The model exceeded expectations, resulting in a mean average precision of **0.969** and **0.980** after post-processing. Next, the model predictions were used to crop out the stamps on each page, to subsequently classify them.

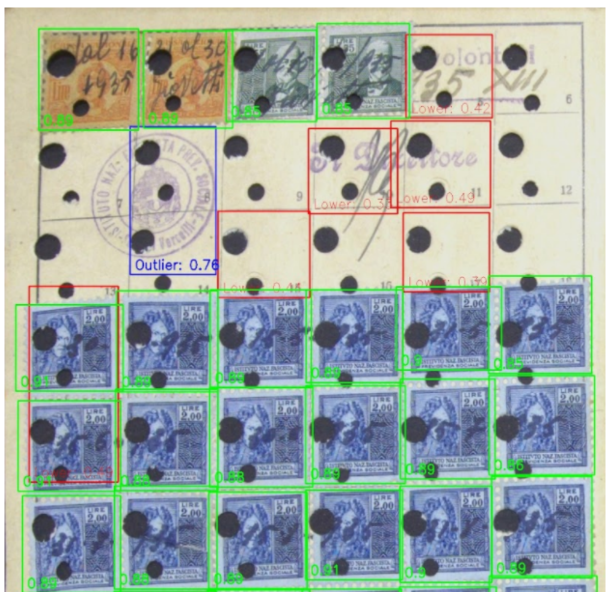

Fig. 2. Output of the YOLO model and post-processing approach, where each prediction is annotated with its confidence score. Predictions marked as correct are shown in green, low-confidence predictions and outliers that were removed in post-processing are shown in red and blue, respectively.

3 STAMP CLASSIFICATION

Because labeling all of the individual stamps is very time-consuming, we developed a custom labeling tool, based on the visual similarity model from [7], to cluster similar stamps together. This sped up the labeling process, by labeling multiple stamps simultaneously, instead of one by one. Each stamp was given two labels, one for the stamp type (the illustration) and one for its monetary value. This is an important distinction because many stamp types had multiple color variants that denoted different monetary values. Some stamps were later adjusted over the years. These adjusted versions had an additional stamp on top, to denote their new value. Most of the stamps were also punctured and had the payout date stamped on top, making them harder to distinguish. A total of 4440 stamps were labeled. Some examples of different stamp types are shown in Figure 3.

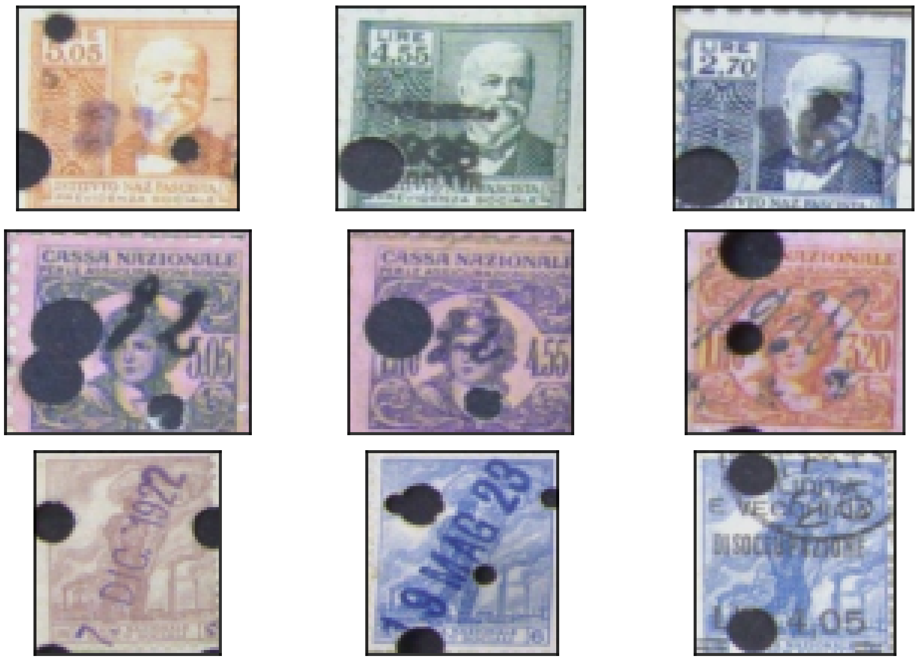

Fig. 3. Each row shows stamps with the same type, but a different value, except for the last image in the bottom row, which denotes an adjusted version, which we consider as a different type. The original stamp has a value of 6 lira, the adjusted one 4.05 lira (denoted by the black stamp on top).

Three different classification approaches were tried. Predicting both the stamp type and value directly, predicting only the value, and predicting both type and value through image similarity. For the first two approaches, we performed transfer learning with a pretrained EfficientNetB0 [9] model. The similarity-based approach used the same similarity model as before. The dataset exhibited a large class imbalance, with some stamps having over 1000 occurrences, while others occurred only a few times. Due to this imbalance, only stamps that occurred at least ten times were considered. For the majority classes, a random sample of 200 stamps was taken. This resulted in 32 classes for the type and value predictions and 12 classes for the type predictions. For the direct classification approaches, a stratified sample of 20% of the training data was used for validation. The similarity-based approach was validated using 5-fold cross-validation. In each fold, we predicted the type and value of stamps in the validation split using the top-1 & top-5 best matches from the training set. The top-5 best matches were averaged to produce the prediction. Table 1 lists the results of each approach.

The results clearly show an excellent performance of the EfficientNet model in predicting just the stamp type, with an F_1 score of 0.94. The same model, trained on both the stamp type and value, performed worse, with an F_1 score of only 0.68. This is mainly because similar stamps had a different monetary value and because of the

Approach	Model	F_1	Accuracy
Type	EfficientNetB0	0.94	0.94
Type + Value	EfficientNetB0	0.68	0.70
Type + Value	Similarity (Top-1)	0.84	0.85
Type + Value	Similarity (Top-5)	**0.84**	**0.86**

Table 1. Classification results for each approach and model

imbalanced dataset. Our similarity-based approach showed an improvement over the classification model, with an F_1 score of 0.84. This improvement can mainly be attributed to an increase in performance on the minority classes, as their most similar matches were often correct. This makes a similarity-based approach a valuable alternative when dealing with large class imbalances on smaller datasets.

This work presented a semi-automatic enrichment pipeline for stamp cards from the 1920s. After preprocessing, we have labeled a small portion of the dataset for stamp detection and classification. A state-of-the-art object detection model was trained and validated on the dataset. With additional post-processing, this model produced impressive results. After extracting the stamps, they were classified using three different approaches. We were able to confidently predict the stamp type and used a similarity-based approach to predict both the stamp type and value. The developed pipeline can now be applied to the full dataset, to provide valuable metadata and open up further research opportunities on the Italian working class of the interwar period.

ACKNOWLEDGMENTS

This research was funded by the AI4EU consortium with the support of the European Commission under the H2020 program within the project DIEA.

REFERENCES

[1] Stephen Broadberry, Claire Giordano, and Francesco Zollino. 2011. A Sectoral Analysis Of Italy'S Development, 1861-2011. *Banca D'Italia. Quaderni Di Storia Economica* 11 (October 2011).

[2] Vittorio Daniele and Paolo Malanima. 2007. Il prodotto delle regioni e il divario Nord-Sud in Italia (1861-2004). *Rivista di Politica Economica* XCVII (2007), 1–49.

[3] Vittorio Daniele and Paolo Malanima. 2009. Labour Supply in Italy 1861-2001 Structural Change and Regional Disparities. *Quaderni Istituto di studi sulle società del mediterraneo* 141 (2009).

[4] Emanuele Felice. 2007. *Divari regionali e intervento pubblico.* Il Mulino, Bologna.

[5] Emanuele Felice. 2007. I divari regionali in Italia sulla base degli indicatori sociali (1871-2001). *Rivista di Politica Economica* XCVII (2007), 359–406.

[6] Claire Giordano and Ferdinando Giugliano. 2015. A Tale of Two Fascisms: Labour Productivity Growth and Competition Policy in Italy, 1911–1951. *Explorations in Economic History* 55 (2015), 25–38.

[7] Filip Radenović, Giorgos Tolias, and Ondřej Chum. 2018. Fine-tuning CNN image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence* 41, 7 (2018), 1655–1668.

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 779–788.

[9] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*. PMLR, 6105–6114.

[10] Vera Zamagni. 1975. La dinamica dei salari nel settore industriale, 1921-1939. 10, 29/30 (1975), 530–549.

[11] Vera Zamagni. 1987. A Century of Change: Trends in the Composition of the Italian Labour-Force, 1881-1981. *Historical Social Research* 44 (1987), 36–97.

[12] Vera Zamagni. 1994. Una Ricostruzione Dell'Andamento Mensile Dei Salari Industriali E Dell'Occupazione, 1919-39. *Ricerche Per La Storia Della Banca D'Italia* V (1994), 349–378.