Wrocław University
of Science and Technology

Faculty of Computer Science and Management
Field of study: **APPLIED COMPUTER SCIENCE**
Specialty: DATA SCIENCE

Master's Thesis

DEEP LEARNING IN COLLECTABLES RECOGNITION

Jan Sieradzki

keywords:
Coin recognition, Deep learning, EfficientNet, SIFT, Coin's bust recognition, Coin's country recognition

short summary:
This thesis treats about collectable coins recognition, with usage of modern deep learning approach. In particular, CNN EfficientNet-b2 is used in purpose to recognize the king's bust minted on the obverse side of the coin. The coins belong to the dataset collected specifically for purposes of this work. The proposed combined EfficientNet-b2 model incorporates domain knowledge about coin's country origin extracted from reverse side, what enhance the obverse king's bust recognition task. During experiments, this proposed method outperforms SIFT + Bag of Visual Words local features based technique and others EfficientNet-b2 based models, which do not incorporate domain knowledge about countries.

Supervisor	D. Eng. Stanisław Saganowski
	Title/degree/name and surname

The final evaluation of the thesis

Chairman of the Diploma Examination Committee	.....	.....	.....
	Title/degree/name and surname	grade	signature

For the purposes of archival thesis qualified to:*
a) category A (perpetual files)
b) category BE 50 (subject to expertise after 50 years)
* Delete as appropriate

stamp of the faculty

Wrocław 2021

Contents

Abstract

This thesis treats about collectable coins recognition, with usage of modern deep learning approach. In particular, CNN EfficientNet-b2 is used in purpose to recognize the king's bust minted on the obverse side of the coin. The coins belong to the dataset collected specifically for purposes of this work. In particular, the dataset, created from images obtained from Polish numismatic online auctions, consists from Russian, Polish, Austrian, German and Roman Empire's historical coins from ancient, medieval and early modern ages. The proposed combined EfficientNet-b2 model incorporates domain knowledge about coin's country origin extracted from reverse side, what enhance the obverse king's bust recognition task by limiting candidates search space to previously predicted country. During experiments, this proposed method outperforms SIFT + Bag of Visual Words local features based technique and others EfficientNet-b2 based models, which do not incorporate domain knowledge about countries. Moreover, CNN's gradient analyze has been performed, and it showed, that deep learning recognition methods are able to find some characteristic coin's regions, which potentially could be used by human in order to recognize coin's type.

Streszczenie

Niniejsza praca dotyczy rozpoznawania monet kolekcjonerskich z wykorzystaniem nowoczesnego podejścia głębokiego uczenia. W szczególności, CNN EfficientNet-b2 jest wykorzystywana do rozpoznawania popiersia króla wybitego na awersie monety. Monety należą do zbioru danych zebranych specjalnie na potrzeby tej pracy. Zbiór ten, stworzony na podstawie obrazów uzyskanych z polskich aukcji internetowych, składa się z rosyjskich, polskich, austriackich, niemieckich i rzymskich monet historycznych z okresu starożytności, średniowiecza i wczesnej nowożytności. Proponowana architektura, bazująca na sieci konwolucyjnej EfficientNet-b2, korzysta z wiedzy dziedzinowej o kraju pochodzenia monety pozyskanej z jej rewersu, co poprawia efektywność rozpoznawania popiersia króla wybitego na awersie, poprzez ograniczenie przestrzeni poszukiwań kandydatów do wcześniej przewidzianego kraju. Podczas eksperymentów, proponowana metoda przewyższa technikę opartą na lokalnych cechach SIFT + Bag of Visual Words oraz inne modele, oparte na sieci EfficientNet-b2, lecz nie wykorzystujących wiedzy dziedzinowej o krajach. Ponadto, przeprowadzona została analiza gradientu CNN, która wykazała, że metody rozpoznawania, bazujące na głębokim uczeniu, są w stanie znaleźć charakterystyczne regiony monet, które potencjalnie mogą być wykorzystane także przez człowieka do rozpoznania rodzaju monety kolekcjonerskich.

1. Introduction

People have always liked to collect various items. They are often devoting great amounts of time and money creating and developing their collection. There exists even the distinct area of study about this phenomenon - psychology of collecting - which explains the emotional background of human desire to gather things. Although people can collect practically everything, there are some objects, which are the most common in this activity. In particular, the coins are among the most collectable objects in the world, probably everyone knows at least one person, who posses a collection of historical coins. As usual, behind the popularity come huge business and also certain problems to solve. One of the main issues in the coin market is preventing illegal trades and coins thefts. Because of this, coins identification is necessary, which traditionally relays on manual searching catalogues of coin in internet auctions, what in practise appears to be ineffective and simply impossible, since the coin market is very active - exemplary, only in the North American market, there are traded half million coins annually [8]. Beyond this, manual classification requires time and expertise due to its complexity, especially for rare, uncommon coins from private or museum collections.

In order to respond to these issues, during last years there were multiple attempts to automatize coin recognition basing on delivered coins photos. Successful development of automated recognition would allow to monitor the market, what implies easy and efficient spotting illegal transactions. Also, for average collector it would be a fast, cheap and comfortable alternative for using, often chargeable, service from physical numismatic expert. The coins are particularly interesting collectable objects to recognize, because despite, that there is a huge amount of coin types, generally coins are similar-looking round objects, what can be described as an interesting fine-grained classification problem. Moreover, coins often contains a lot of cultural, historical and artistic elements minted on their both obverse and reverse sides, which often are used to coin recognition. Especially in the last years, new approaches are rather trying to understand minted motifs context and type of the coin, than simply to classify them by comparing with other specimens in database. Over the past years, multiple techniques were used in order to automatize coin recognition. Recently, due to dynamic development of deep learning, methods based on Convolutional Neural Networks (CNNs) are in the center of attention. Many CNN models have been created and tested on the Ancient coins, which are by far the most popular type of coins in recognition researches. For this reason, this works decided to change the collectable coin's subset and perform researches on less popular coins, mainly medieval, which frequently occur in Polish numismatic trade market. In results, success of deep learning methods obtained on the dataset consisting from coins taken from real-life auctions, gives hope, to potential application of such solutions to the real trade market (in this case to Polish one) and finally prevent, previously mentioned problems, like illegal trades and coins thefts in the internet auctions.

Aim of the work

The aim of this work, is to perform recognition of the ruler's head/bust minted on the obverse side of the coin, with usage of a modern deep learning method and domain knowledge incorporation. In particular, the dataset used to evaluation, should contain coins with the ruler's bust/head minted on the obverse side, which are representation of the most common collectable coins types occurring in the Polish numismatic trade market, including coins from different periods of time and regions. In order to keep proper difficulty and diversity level, dataset should consist mostly from non-modern coins. Proposed method is expected to outperform local feature based method SIFT+BoVW from paper [2] and simple deep learning method without domain knowledge incorporation.

Research Questions

The research conducted in this thesis, will lead to answering the following questions:

- Which modern deep learning method is appropriate for obverse's bust/head recognition for collectable coins dataset?
- Will deep learning based methods outperform local feature based methods in recognition of diverse coins types from Polish numismatic trade market, like it has already happened in the case of Ancient Roman coins datasets [1, 8, 14]?
- What domain knowledge can be incorporated to deep learning model in order to enhance obverse's bust/head recognition?
- Will incorporation of the domain knowledge enhance recognition capability of the deep learning model, or rather make it worse, as a deep learning model itself is able to extract own features and perform better?
- Will proposed deep learning method be capable to understand semantic context of the coin and recognize obverse's bust/head with usage of elements, which would be in the field of interest for manual classification performed by human?

Scope of the work

In order to fulfill the aim of the work and answer for addressed research questions, following works have been planned to carry out. The literature in scope of coins recognition, convolutional neural networks and local features extraction should be reviewed and summarized. A balanced and labeled dataset need to be collected, which will be including medieval/early modern coins from the most common coins types occurring on the Polish numismatic market. There should be proposed method to recognize the head of a ruler minted on the obverse of coins from the collected dataset. The method is expected to use a modern deep learning techniques and domain knowledge about the rulers minted on the coins. Deep learning and local feature based methods are planned to be implemented and compared in the experiments. Finally, the experiments should be carefully analyzed, with the help of traditional measures for classification task, Grad-CAM CNN model's gradient visualizations and Friedman test with Shaffer's post-hoc in order to check statistical significance of obtained results. All mentioned steps will be described in this thesis in details.

Chapters content

The rest of this thesis is organized as follows. Chapter 2 explains the coin recognition background, including coin classification methodology, definitions of important terms (section 2.1), related works, existing methods (section 2.2) and presentation of the coin dataset, collected on purposes of this work (section 2.3). Chapter 3 describes methods, which have been used during experiments. In particular, models based on convolutional neural network family EfficientNet [17], including proposed combined model, are introduced (section 3.1) and local feature approach SIFT+BoVW+SVM [2] (section 3.2), which has been chosen as the baseline method for considered dataset. In chapter 4 there are presented results and analyze of performed experiments (section 4.2), with aim and methodology explanation beforehand (section 4.1). Finally, chapter 5 summarizes carried out works with evaluation of the objective, conclusions and proposition of future work in this area. The paper is finished with the list of literature used in this thesis.

2. Problem analysis

In this chapter there are described methodology and main challenges of collectable coins recognition. There are presented main ideas, which have been used to create, proposed in this work, approach to automatic coin recognition. Then, there is presented an overview of related works and methods. Finally, the dataset, collected for purposes of this work, is described and analyzed.

2.1 Coin recognition challenges

Coins are very interesting objects in terms of recognition, from economic reasons, as they have always been the dominant currency in human history, as well as from academic reasons, because of the presence of archaeological, artistic, historic and cultural symbols on both sides of circular-shaped objects, which usually are similarly-sized. As coins are from definition naturally valuable trading items, the need of reliable recognition is obvious in order to avoid thefts and scams on unaware holders.

2.1.1 Coin's characteristic

In purpose to recognize coin type, there is necessity to distinguish characteristic components of this object. In fact, in numismatic certain parts of the coin are distinguished and can be found in the majority of coins. Figure 2.1 presents these characteristic coin's parts on an example of Ancient Roman coin [3].

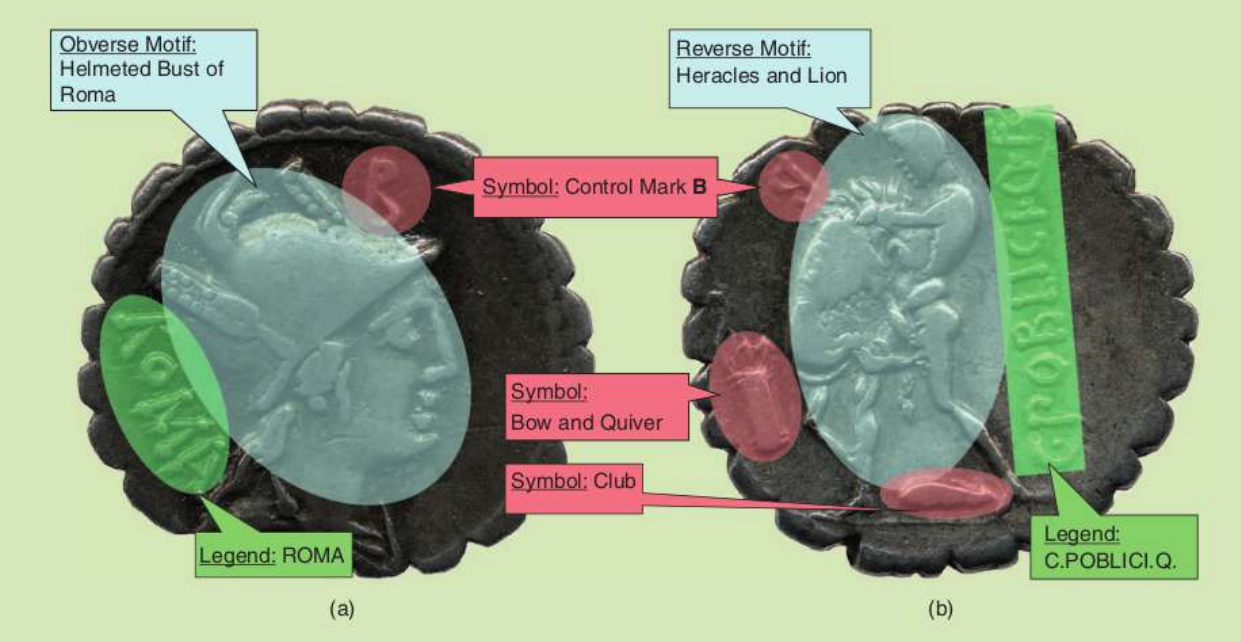


Figure 2.1: Basic coin elements on the obverse (a) and reverse (b) on example of Ancient Roman coin, graphic source: [3].

In order to classify coin type, a numismatic expert may analyse the following coin's attributes:

- Physical characteristics
- Obverse legend
- Obverse motif
- Reverse legend
- Reverse motif
- Minor symbols

On the both sides of coins, there can be distinguished main motif, legend and symbols. The head of a coin (obverse side) usually contains a portrait of the actual ruler (issuer of the coin) as the main motif. In the other hand, the reverse side's main motif is more diverse, there are common motifs types like blazons, weapons, saints, Gods (in Ancient Rome), places, eagles or other animals. The legends, often present on both sides, usually contains the descriptions of main motifs, including name of ruler present on obverse and year of his reign. Information about alphabet used in legend delivers very discriminate clue, what significantly reduces search space - this information is perfect for classification and is natural to use, but unfortunately the legends are the most prone coin's elements to damages and erase, hence recognition can not relay on this part in case of old, often significantly worn, coins. Moreover, the legend often may be not understandable for average, non-expert, collector, because of alphabet difference, therefore automatic recognition still would be useful. Additionally, there are minor images, often used as mint markings, like crests, clubs or numerals, which in numismatic terminology are referred to as symbols. Finally, there are coins physical characteristics - weight, die alignment, diameter, colour, specific coin's tears - which are often used by mechanical devices to automatic coin recognition/sorting purposes and by experts for manual recognition.

2.1.2 Coins intra-class variability

In the past, coin minting was not as consistent, as nowadays. There can be observed many medieval/ancient coin classes with small count of specimens for some of them, hence, there is added to the recognition task new rarity challenge to overcome. Models need to classify coins, which were not present in the training phase - in this case accuracy is especially important, since rarity often means higher item's prize. Coins of the same type (class) have the same elements, e.g. the bust of a ruler which has some particular clothing (crown, laureate, cuirass ...), the same reverse motif and the same minor symbols. Semantically, in the single class, the elements will be the same, however their depictions can significantly differ, especially in the case of ancient and, in a smaller degree, in the case of medieval coins (exemplary figure 2.2). The big differences in appearance variability are caused by the following facts:

- Same class coins could be minted with usage of die created by different engravers, causing serious dissimilarities
- The other variances could be created by wear and tears of a coin (due to its age or conditions in which it was preserved), metal's surface could even change the colour due to oxidation
- Differences in coin's material
- Wear and damages of dies used to coin mint
- Different die centring - in the mints from the past, die centre not always have coincided with the piece of metal used to coin creation, the poor centring can cause missing of elements close to edges, e.g. part of legend, (Neron on figure 2.2)

These factors are having big influence on appearance of coins and their high intra-class variability, what beyond making them more collecting-attractive and expensive, creates serious obstacles in manual and especially automatic coin classification. That is why modern coins datasets are much easier challenge for visual processing methods and despite, that there already exist high performance solutions regarding this type of coins, these successes do not fully transfer to much more demanding ancient/medieval coins and there is still a lot to improve in that field.

John III Sobieski, Poland

Aurelian, Roman Empire

Catherine II, Russia

Sigismund III, Poland

Neron, Roman Empire

Figure 2.2: Intra-class variability examples from this work's dataset.

2.1.3 Domain knowledge incorporation

Considering high intra-class variability, it is logic to incorporate domain knowledge into models, to make them understand which artistic concept is located on the given coin element - in other words, make the model work in the way, a human expert would classify coin, identifying the individual semantic elements depicted there on [4]. This approach would allow the model to recognize far more coin types, even if these types would not be present in the training dataset. The usage of domain knowledge, about obverse and reverse elements, is in the field of interest in the recent automatic coin classification researches [1, 2, 3, 4, 8, 14] e.g. usage of knowledge about obverse, that there is located bust of coin issuer or about legend's alphabet, which delivers information about the coin's region, or about reverse motif association with country's culture (the last idea is used in this work).

Overall, coin classification can be defined as fine-grained classification, as the objects belong to one similar-looking super class `coin` and need to be distinguished on subclass. Usually, a subclass is equal to the ruler from an obverse's portrait who issued the coin - it allows extracting information about country, time and year of coin minting. Some other time, subclass will be described with more discriminative combination of obverse portrait and reverse motif [8]. Nevertheless, a lot of approaches focus on limiting the search space, instead of finding exact subclass, getting closer in direction of coarse-grained classification. Papers [2, 1, 3] are focusing on recognizing reverse motifs, which are big elements with high-variance of shapes and meanings. As the same reverse motif can occur on many classes (in particular on coins with different obverse portraits), this task can be described as coarse-grained classification.

2.1.4 Coin's motifs seen by CNN

In general, as the main motifs are the biggest coin's elements, and hence are more resistant against degree of wear, they are the part of interest in multiple approaches, which perform coarse-grained coin classification (are focused on recognizing big components present on a coin, which may be associated with multiple coin's types - reducing search space instead of searching for direct coin's class). In the paper [8], the authors have visualized CNN, trained on both obverse and reverse sides images, and in result have obtained class-sensitive regions on the coins. These visualizations have shown, that deep learning model's discovered regions are consistent with human expert annotations. Furthermore, the obverse images needed more regions to remain distinguishable, than the reverse images - it results from the fact, that all obverse images contain approximately the same shape (face or bust). In contrast, on the reverse side motif there are presented very heterogeneous, hence easier to distinguish, objects. It is worth to note, that obverse's motif similarity implies necessity of analysing obverse details, which will often be erased in discussed ancient/medieval dataset (hence motif distinguishing becomes hard). That fact of smaller inter-class variation makes reverse motif recognition easier, but in the same time, obverse motif delivers more interesting information, in context of coin classification, than reverse's one (the issuer name, inferred from the obverse motif, is more class discriminative, than the reverse motif). Hence, good idea is to use reverse motif in purpose to conduct coarse-grained classification and later refine such reduced search space by fine-grained classification (classification of very similar objects in purpose to find out specific class), exemplary basing on obverse portraits [8].

2.1.5 Similarity of face and obverse recognition tasks

The task of obverse's head recognition can seem to be similar to face recognition task, which has been deeply explored during decades and is dynamically developing in terms of deep learning usage during last years. In fact, both cases are sharing intra-class appearance variation due to facial hair, clothing or age. On the other hand, the illumination, colour and edges features are not as discriminative in the case of coins, as in the case of real faces, because the coin's material and obverse details can differ within a class. Coin ruler will be recognized more often by characteristic details, like hairs style, crown or clothes parts from the bust. However, as are showing CNN's visualizations from [14], for coins created in more detailed, artistic, and expressive manner, the actual facial part of a head are sufficient for successful classification. Furthermore, the head on the coins is usually presented in the profile, what is more challenging than frontal or semi-frontal poses, which are more frequently the case in the face recognition studies.

2.2 Related work

Over the past years, researchers have proposed multiple approaches to implement image-based coin classification framework. In this section, there is an overview of the related work done to date in the collectable coin's recognition area.

2.2.1 Modern coins recognition

The earlier works were focused on modern coins [12], as these, due to modern methods of manufacturing, present small intra-class variability (for computer vision two coins of the same type are practically identical), hence are much easier to recognize, than medieval or ancients ones. This allows usage of straightforward holistic methods based on edges features, like in [19], but the good recognition quality can not be transferred to medieval or ancient coins due to their edge's variability. Another paper [12] utilizes direction of the coin's gradient vectors. In order to align two compared coins, the authors search for optimal rotation, which maximize the correlation between corresponding coin's gradient vectors direction. Then they classify them with nearest neighbour search, supported with several rejection criteria. Meanwhile, this method is proved to be sufficient for easy tasks, like sorting modern currency, it is not capable to work with challenging conditions of older coins. Moreover it will not recognize the type of coin without seeing it beforehand, what is desirable in case of rare coins. In fact, in [20] authors showed, that edges [19] and gradient [12] techniques success with nowadays-used coins classification does not transfer successfully to classification task of ancient coins.

2.2.2 Methods based on local features

In order to recognize types of coins, which were not previously seen by model, papers [3] and [2] propose an approach similar to human intuition - incorporate domain knowledge, by making model able to recognize characteristic motifs minted on the coin i.e. make the model work in coarse-grained classification way, recognizing big components on the coin, which can be minted on more than one coin class, where coin class is understood as a coin with the obverse and reverse motifs of the same type. In [2] authors are using local features SIFT [10] and Bag of visual words (BoVWs) - the concept adapted from information retrieval and NLP, which found the application in computer vision problems, like object and scene recognition. Specifically, authors of [2] utilize this technique on three reverse motifs for ancient coin's coarse-grained classification, claiming that reverse motifs are more discriminating, than the portraits of the emperors on the obverse side (emperors faces are more similar to each other, than reverse motifs, which with high probability will have entirely different shape and structure). Moreover, they solved BovW's issue with the lack of spatial information, by using circular tiling. Circular tiling outperformed rectangular tiling and log-polar tiling in their experiments and at once circular tiling maintains rotation invariant, which is lost in case of the other methods. Nevertheless, lack of spatial information still remains as a main drawback of SIFT based technique, alongside with computational heaviness. Additionally, authors of [14, 4] notice, that circular tiling technique assumes, that coins have perfect circular shape and centring, what is not really realistic, especially in the case of Ancient coins. Work [3] enrich circular tiling with geometric visual words relative relationships, but in experiments there is only 1% point accuracy gain in comparison to circular tiling alone. As the method from [2] has been chosen as a benchmark in this work for a proposed deep learning method, it is explained in details in section 3.2.

2.2.3 Hierarchical recognition

The authors of work [8] used, pre-trained on ImageNet, CNN AlexNet [9] model to three tasks: reverse recognition, obverse recognition, and hierarchical reverse recognition. As the first two tasks are classical problems, the third one presents an interesting approach, which improved reverse recognition by 22%. The idea behind this method is to train two models, one for reverse and the second for obverse recognition. Then, after calculating the probabilities for two classic tasks $p(emperor_i|I_{obverse})$, $p(reverse_i|I_{reverse})$, probabilities should be combined for every possible emperor/reverse shape combination. Additionally, there are considered only emperor/reverse reasonable combinations given by *a priori* knowledge, what is fulfilled with adding to formula following factor: $\delta(Pa(reverse_i) = emperor_i)$ where $\delta()$ is the indicator function and $Pa(reverse_i) =$

$emperor_i$ means that $reverse_i$ has $emperor_i$ on the other side. Finally, there will be returned the predicted reverse's shape which maximize probability:

$$p(emperor_i|I_{obverse_i}) \cdot p(reverse_i|I_{reverse}) \cdot \delta(Pa(reverse_i) = emperor_i) \tag{2.1}$$

This method is presented in tree structure on figure 2.3.

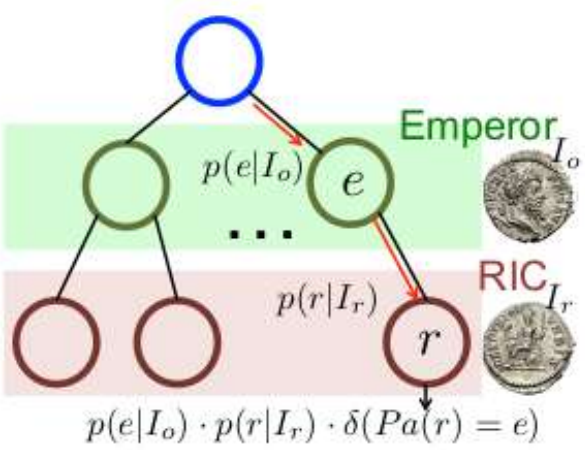


Figure 2.3: Hierarchical classification of obverse, graphic source: [8].

Summing up, hierarchical functions combines two probabilities and takes, as a result, the most probable option, not taking into consideration combinations which, according to certain *a priori* knowledge, should not appear together. It is worth to note there, that this combination's exclusion, although certainly prunes some not interesting pairs from further classification and improves performance in this way, in some cases can eliminate a solution from the considered set of candidates as a result of a failed obverse (emperor) classification. Thus, sometimes it can propagate error, instead of useful information. Nevertheless, it is still a very interesting idea to use emperors as a coarse grained classification and afterwards, in the obtained search space, look for the reverse shape, which fits best to the predicted obverse. The combined CNN model proposed in this work (more in section 3.1.3), which supports obverse's ruler recognition with country classification performed on reverse, has been inspired by this hierarchical idea.

2.2.4 Methods based on deep learning

There was already mentioned, in section about local features based method 2.2.2, that researchers in [3, 2] tried to incorporate domain knowledge into automatic recognition, in order to make automatic classification human-similar and make it understand, what context actually is presented on the coin. This idea allows getting closer in a very interesting and useful direction of recognition of coin's overall type, instead of recognition of specific coin. Nonetheless, technological limitations of local features approaches like `SURF` or `SIFT`, especially lack of object's spatial analyse (despite the attempts to minimize this drawback with circular tiling [2] or key point's relative position geometric [3]), makes these methods coins 'understanding' ability rather disappointing [14].

In paper [8], already described in the hierarchical recognition context (section 2.2.3), CNN architecture model, AlexNet [9], was used, which at the moment of release (2012) was state of the art on ImageNet dataset classification and one of the pioneers of deep learning CNNs. However, at the time of writing this work, it is strongly outperformed by much newer conceptions. The ImageNet is a large database containing more than 14 millions annotated images, designed for use in visual object recognition research. In [8] authors decided to fine-tune CNN model pre-trained on ImageNet, as the AlexNet architecture is large (60 millions parameters) and their dataset (about 4500 images) is too small to avoid under-fitting in case of training from scratch. ImageNet fine-tuned deep learning model allowed to overcome the limitation of the small data and the model

managed to significantly outperform traditional approach based on SIFT local features and SVM classifier. In particular, CNN's model obtained 45.6% better performance in reverse motif recognition, 21.2% better in obverse recognition and 25.5% better in hierarchical obverse recognition, what can be considered as a huge improvement and suggested potential deep learning dominance in coin recognition problem. The especially big performance gain for reverse motif recognition reason is, as authors of [8] claim, the CNN's ability to effectively exploit the spatial information, in comparison to SIFT local futures, which lacks such ability. Particularly, as in the reverse side of coin the spatial information is crucial in the face of big variability of symbol's shape and structure, the CNN's advantage was in particular visible in this task.

Authors of [14] decided to perform emperors recognition on Ancient Rome coins. In that purpose they used own-crafted net, constructed from 5 convolution blocks, inspired by Simonyan and Zisserman architecture from paper [16], where the main idea is to use small (3x3) stacked kernels, instead of the big one. Their method is shown to outperform SIFT approach by an order of magnitude. As their CNN base architecture [16] was released in 2014, it is logical to believe, that modern architectures can even improve this gap.

In paper [4], which was focused rather on interpreting and visualization of CNN's, instead of obtaining the highest possible accuracy, the architecture loosely based on AlexNet [9] was used to classify reverse objects on huge Ancient coins dataset. AlexNet, as it was already mentioned, is not competitive anymore since years, nevertheless paper [4] still made important contributions in field of deep learning in coins. The authors performed unstructured text analysis of coins automatically taken from internet auctions. In that way, they created weak associations between semantic elements found on ancient coins and their images, and made AlexNet learn the appearance of five reverse elements (horse, cornucopia, eagle, patera, shield). It occurred, that their model successfully was finding aforementioned elements, despite the huge differences between coins (one element is present on multiples classes), what would be impossible without semantic understanding of coin. Thus, in work [4], in practical, experimental way, there is shown the advantage of context understanding over straightforward direct coin's comparison.

Work [1] presents CoinNet, the authorial deep learning architecture, which is build with two popular CNNs - DenseNet161 [7] and ResNet50 [6] - enriched with attention mechanism. Authors state, that they gathered the largest Roman Republican coins (18000 images, 228 different reverse classes). They performed reverse recognition and have obtained more than 98% accuracy with proposed CoinNet, outperforming other, recent state-of-the-art deep learning methods. They claim, that wrong classifies were often caused by small image's resolution, what is interesting observation. Paper [1] proved, that with availability of huge data, state of the art CNN can assure very robust classification. However, it is still the open question for a smaller dataset, like presented in this work, which includes coin from other regions and ages. Moreover, the authors of [1] were focused on reverse recognition, unlike in this work, where the main goal is to recognize the bust presented on the obverse side.

2.2.5 Literature review summary

Literature, in the area of coin recognition problem discussed in this work, is mainly focused on Ancient coins problem, as previous subject of researches, modern coins, can be considered as solved and not demanding enough. There is no a lot of works, which deal at once with ancient, medieval and early modern coins, what is the case in this thesis. Furthermore, because most researches concern only Roman coins, there is no discuss about recognizing the coin's origin country - rather there are attempts to understand semantic context of reverse motif minted on coins. Regarding methods used to coin recognition, previous local feature extraction methods, like SIFT, used to be the state of the art, but during recent years deep learning approaches started to significantly outperform other methods. However, traditional methods, like edges or local features comparisons are having important advantage, that their success does not depend on from abundant datasets, like it is in the case of CNN's models. This creates problems, especially for rare coins, which will not appear during training. Fortunately, in the case of deep learning it is compensated with good understanding of coin symbols, what allows the model to perform well even for zero-shot classification. On the other hand, when there is a lot of data available, deep learning approach shines, whereas local features method becomes very computationally heavy [1]. Taking everything in account, nowadays, when deep learning becomes more and more efficient with every year and there are already available light-weight models offering impressing accuracy for complicated problems, deep learning can be considered as far best choice in the recognition task. Furthermore, considering deep learning incredible ability to understand coin's context, it seems that CNN's

solution definitely outperform traditional methods, thus in this work there is proposed the new deep learning approach. Method presented in this work profits from hierarchical recognition idea taken from [8] (section 2.2.3) and strength of CNN's from ImageNet state of the art's family EfficientNet [17], in light-weighted version EfficientNet-b2 proposed in 2019. The method will be used on the new, diverse, dataset, collected specifically for the purpose of this work, in order to recognize the ruler present on the coin's obverse, with usage of information extracted from reverse side. In particular, information taken from reverse side can be called domain knowledge incorporation, as there is used ruler-country association and the fact, that coins from each country are having some characteristic appearances and motifs. Because of that, it is reasonable to believe, that country recognition can be performed easily and successfully enhance obverse's classification, by limiting potential kings search space to rulers from a predicted country.

2.3 Dataset

Large majority of the researches on collector coins are focused either on Republican Roman or on Roman Imperial coins. The earlier works in that field also were treating about modern coins, but this problem can be considered as solved due to high objects quality and small variance. The reason, why Ancient Roman coins are in the highlight, is that in fact are the most interesting and challenging currency, because the characteristics mentioned in previous sections, like huge amount of classes or big intra-class variance. Thus, during the years of researches, there have been created a lot of great datasets with Roman coins, which were repeatedly analyzed. On the other hand, it is not easy to find dataset with medieval coins and even if, usually are not ideal for recognition research purposes due to weak annotation and no class-balances. Coins, from medieval time and even early modern age are having similar issues, thus interesting characteristic for researches, like in the case of Ancient Roman coins. These coins also are marked by high variability, amount of class, wears, manufacture defects (different dies used to minting, not ideal centering), thought probably in a bit smaller degree, than their Ancient Roman equivalents. However, there are very visible differences between coins from various countries and time periods, thus it can be interesting to include in dataset different types of coin, instead of working only on a one super-class (like moderns or Ancient Roman coins), what usually was the case in foregoing papers.

In this work, the dataset consists from 1956 coins (look table 2.1) from four different medieval European countries: Poland, Austria, Russia, Germany and from ancient Roman Empire. An important attribute of every coin in the dataset, is having on the obverse side head or bust of the issuer - i.e. actual ruler of the realm. For each country there have been chosen 6 rulers. On the figures 2.5 and 2.4 it is possible to observe distribution of coins in the country and ruler classes respectively. The dataset can be considered as rather balanced, especially in terms of countries, however some rulers classes (Sigismund I Old, Anna Ioannovna) seem to have significantly smaller amount of coins.

Table 2.1: Number of coins in the dataset for each ruler and country.

All coins								1956	
Russia	**340**	**Austria**	**395**	**Germany**	**364**	**Poland**	**430**	**Rome**	**427**
Alexander III	66	Ferdinand II	50	Friedrich Wilhelm I	84	Augustus III the Sas	88	Aurelian	74
Anna Ioannovna	32	Ferdinand III	49	Friedrich Wilhelm II	77	John II Casimir	84	Constantine I	72
Catherine II	62	Franz Joseph I	75	Friedrich Wilhelm III	60	John III Sobieski	77	Hadrian	68
Elizabeth	77	Karl VI	80	Ludwig I	41	Sigismund I Old	25	Neron	65
Nicholas II	58	Leopold I	90	Ludwig II	41	Sigismund III	95	Philip I	70
Peter I	45	Leopold V	51	Maximilian III Jose	61	Stephen Báthory	61	Trajan	78

It is important to note, that inside the ruler class, beyond the coin's conditions diversity, there are usually multiple semantic variants of coins, i.e. various ways of bust's exposition or various types of reverse motif (exemplary samples are presented on images 2.8, 2.9). Another important matter, with influence on the variability, is the age of the coins. Six rulers per country means six different periods of time per country. Table 2.2 presents the years of the beginnings of the reigns for each ruler respectively - this can be associated with approximated minting time of the coins having given ruler's head on the obverse. Figure 2.6 shows the distributions of centuries, in which rulers took the power. It is visible, that most coins were minted approximately in XVIII century. There is also a very clear split on medieval/ early modern coins and Roman Empire cluster.

All images in the dataset are, similarly like in [4] and partly in [1], taken from internet auctions. Specifically,

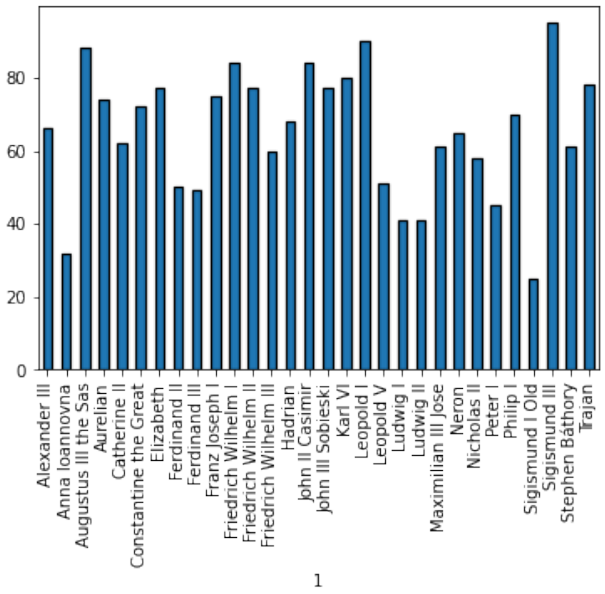


Figure 2.4: Histogram of coins for each ruler.

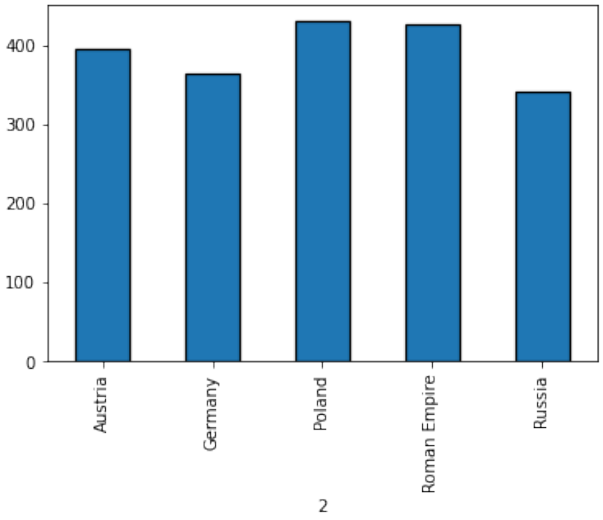


Figure 2.5: Histogram of coins for each country.

Table 2.2: The year of the beginning of the reign for each ruler.

Russia		Austria		Germany		Poland		Roman Empire	
Alexander III	1881	Ferdinand II	1617	Friedrich Wilhelm I	1713	Augustus III the Sas	1733	Aurelian	270
Anna Ioannovna	1730	Ferdinand III	1637	Friedrich Wilhelm II	1786	John II Casimir	1648	Constantine I	306
Catherine II	1762	Franz Joseph I	1867	Friedrich Wilhelm III	1797	John III Sobieski	1674	Hadrian	117
Elizabeth	1741	Karl VI	1711	Ludwig I	1786	Sigismund I Old	1506	Neron	54
Nicholas II	1894	Leopold I	1658	Ludwig II	1864	Sigismund III	1587	Philip I	244
Peter I	1682	Leopold V	1177	Maximilian III Jose	1745	Stephen Báthory	1575	Trajan	78

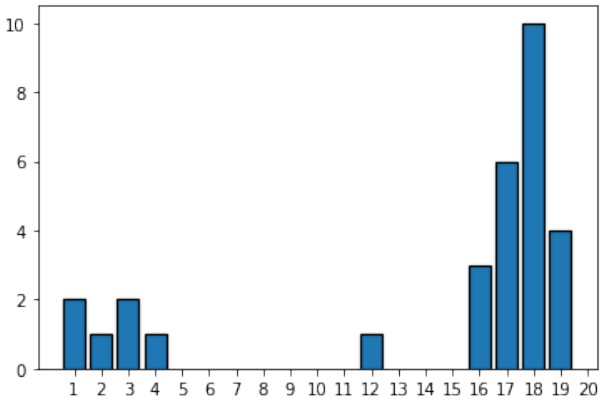


Figure 2.6: Histogram of the centuries, in which the rulers in the dataset began their reigns.

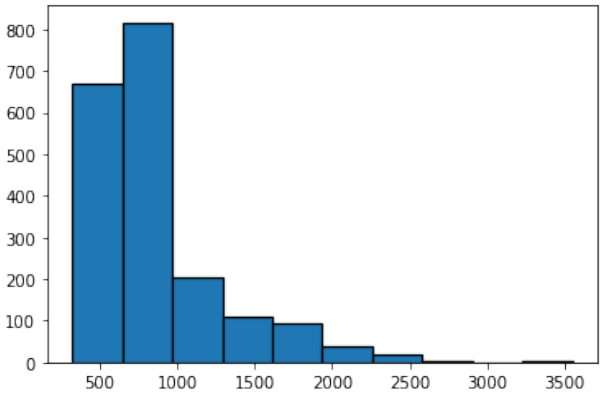


Figure 2.7: Histogram of image's resolution. In particular, image's height is plotted, whereas image's width is approximately equal to $2 \cdot height$.

this work is focused on Polish numismatic market, hence the country and kings classes were not chosen randomly - these classes are representation of the most popular coin types traded in Poland and usually taken right from Polish auctions (medieval Poland shared borders and often was altering its territory with Russia, Germany, Austria. Furthermore, Roman Empire had influence all over the Europe, thus there is a lot of historical coins from these countries in Polish market). Every offer had the same format of images - obverse on the left concatenated with the reverse on the right. In all cases, both coin sides are fitted into a rectangle with white background, what have spared the necessity of coin segmentation and cropping. The dimensions of images differ significantly, however, because of the coin's rectangular shape, every image's height:width

ratio is 1:2 (obverse and reverse images are concatenated horizontally, exemplary images from dataset are visible on figure 2.2). On figure 2.7 there is histogram of image's height ($width = height * 2$). The image in the worst quality has 324 height pixels - the one with the best 3549 pixels. The most frequently, height pixels amount lies in the range [324,969]. The overall quality of dataset's images can be considered as sufficient for recognition task, what has big importance, as authors of paper [1] have experimentally confirmed, that their high quality deep learning architecture failed mostly on the images with low resolution.

Important factor, which in big degree determinants the recognition hardness, is coin's condition. In auctions, sellers usually are putting coin's condition ratings in the description. Unfortunately, beyond the fact, that auction's descriptions are not structured, thus it is difficult to extract particular information from them, sellers are often using different condition's evaluation scales. Moreover, condition evaluation is performed manually and there is always existing human's bias in the particular grades. The estimated coin's states ratings often seem to be overstated, as good coin's condition raises the price, what is in the interest of a seller. The automation of the coin's condition evaluation is certainly an interesting direction of researches, as it would introduce independence of human bias and unify coin condition scale. However, because of mentioned reasons, the overall condition of coins in the dataset can be only approximated by manual review. As the data are sampled from real offers from trade market, the coins are having very wide range of states. There can be found significantly worn specimens and on the other hand specimens in the ideal state. As we can observe on sampled coins on figures 2.8, 2.9, the most of the coins are showing the signs of age and some details are erased, but in general the main motifs, sometimes even legends, are readable and possible to recognize. Figure 2.10 presents reverse motifs sampled from each country. It is worth to notice, that each country has fairly characteristic style of reverse's motif style and context. Russian and Austrian coins often have minted their eagle of arms. Poland commonly has different kinds of shields with minor symbols, meanwhile German and Roman coins presents wide, hence easy to distinguish, spectrum of persons, animals, scenes and text content.

Figure 2.8: Obverses of coins from dataset, random 6 samples for exemplary two rulers from each country (from top rows, two kings from Russia, Austria, Germany, Poland and Roman Empire respectively).

Figure 2.9: Reverses of coins from dataset, random 6 samples for exemplary two rulers from each country (from top rows, two kings from Russia, Austria, Germany, Poland and Roman Empire respectively).

Figure 2.10: Reverses of coins from dataset, random 12 samples for each country present in the dataset.

3. Proposed method

This chapter explains methods, which have been used in experiments to recognize coins from the dataset introduced in this work. On the beginning, there is description of EfficientNet deep learning family methods. In particular, there is explanation of the combined EfficientNet-b2 model, which incorporates domain knowledge about coin's country origin, in order to enhance ruler's bust recognition performed on obverse side of coin. Afterwards, there is description of approach based on local features extraction SIFT+BoVW+SVM [2], which has been chosen as a baseline algorithm for proposed method.

3.1 Combined classification with EfficientNet-b2

In this section, EfficientNet [17] CNN family has been described. Afterwards, combined EfficientNet-b2 model for obverse's coin recognition is introduced and explained.

3.1.1 Convolutional Neural Networks

Deep learning is an artificial intelligence function, which is inspired by the human brain's way of processing data and patterns creation in decision-making. In recent years, deep learning has become one of the most interesting and developed branch of machine learning, due to its very powerful ability to handle big data and context understanding. It is considered as a breakthrough technology, which opens the door to rapid improvements, exemplary regarding self-driving-cars, drones, security, medicine and robotics. In particular, Convolutional Neural Networks (CNN) have grown to be a popular choice for deep neural network.

Convolutional Neural Network is an algorithm, which is taking an image as the input and through learnable weighs and biases can assign importance to certain image's features, what leads to context understanding ability and thus CNN's capability to differentiate one images from the others. The advantage of this method is very low pre-processing requirement, as it is able to learn features by its own through capturing the temporal and spatial dependencies, instead of taking usage of hand-engineered features what is the case in the big part of traditional approaches. The main idea behind CNN's is the image's tensor representation reduction into compact form without lose of features, what allows for easy processing and high quality classification. The heart of the CNN's lays in convolution operation, which relay on moving the filter all around the convoluted tensor, producing a new, reshaped output tensor. Convolution can alter the input's tensor shape with respect to following most commons parameters:

- amount of filters (depth-dimensional size change),
- stride (width and height of tensors 'slices' - spatial change)
- filter size (width and height of tensors 'slices' - spatial change)
- padding (width and height of tensors 'slices' - spatial change)

thought there are many more to tune. Among others important CNN's operation, there are pooling layers, which are responsible for spatial size reduction of previously convolved features - this decreases required computational power necessary for data processing, at once being useful in terms of extracting rotational and positional invariant dominant features. Usually, pooling and convolutional layers are occurring together, creating i-th layers of a Convolutional Neural Network, where the total amount of such layers determines the depth of CNN. In simplify, the deeper CNN is, the more complex features can be extracted, but at the cost of computational power growth. Generally, initial convolution's in CNNs are responsible for extracting the low-level features like colour, edges or orientation of the gradient. Deeper convolution layers, with previous

low-level outputs, are building up high-level features, which are having strong images distinguishing and understanding capabilities. At the end of CNN, usually the tensor is flattened and putted into an attached neural network for a final classification (if CNN is used for classification task).

CNN's, primarily, were used to tasks like image's clustering and classifying or object's recognizing in scenes, but since there is huge boom on this technology, especially after AlexNet [9] success in 2012, CNN has found its applications in many other fields, exemplary in text analytic. As nowadays, the researcher's interest in Convolutional Networks is huge and is still growing, there are constantly released papers, which are presenting better and better architectures and other improvements enabling performance boost. For this reason, it is important to follow trends and implement them for old and known problems, using the latest ideas for result's enhancement. Eventually, this attitude was motivation to take a use from the state of the art's CNN EfficientNet [17] and check it out on the coin dataset considered within this work.

3.1.2 EfficientNet

Authors of [17], in 2019, have made very important contributions to deep learning field. They have proposed a new CNN's scaling method, which scales uniformly all three dimension of convolutions (resolution, width, depth) with usage of simple, compound coefficient. The generalization of this method has been demonstrated on the others, popular-choice CNNs: ResNet and MobileNets. Furthermore, in [17], there has been proposed a new architecture called `EfficientNet-b0`, which is easly-scalable and have become a baseline of EfficientNet family, where subsequent models are obtained from scaled-up baseline. At the moment of writing this work, net `EffNet-L2(SAM)` [5] is the TOP 1 model on CIFAR-10 and CIFAR-100 benchmarks, meanwhile `Meta Pseudo Labels` [11] is the TOP 1 model on ImageNet benchmark - these solutions are highly based on EfficientNets-L2 (successor of EfficientNet-B7 [17], which is scaled, big version of EfficientNet-B0 baseline).

Scaling CNN

Scaling up is an important approach, used to improve net accuracy. Generally, there are three main ways to scale up convolution nets:

- Depth scaling - deeper CNN can capture more complex features and generalize well. Very deep networks can cause some difficulties during training, due to the vanishing gradient, what is minimized by techniques, like skip connections or batch normalization.
- Width scaling - wide networks are able to capture more fine-grained features and at the same time, not make the training much more difficult. But in the case of very wide and shallow model, there appear problems with capturing high-level features. Especially popular for small models.
- Image resolution scaling - high resolution images contain a lot of details, what allows model to extract more fine-grained patterns.

However, in the past usually only one of the mentioned dimensions (depth, width, image size) was common to scale at the time, because scaling two or three of them required difficult manual tuning, what not always resulted in performance improvement. It has been empirically observed, that scaling dimensions are not independent (if on the input there is big sized image, it is intuitive, that network should be deep in purpose to process all pixels and extract high-level features). Moreover, scaling up only one dimension improves accuracy only until a certain moment, later accuracy gain diminishes for big models. Taking that all into account, authors of [17] proposed to scale simultaneously all of three dimensions (visualization is visible on figure 3.1) with a set of fixed scaling coefficients.

In the equations 3.1, there are d,w,r parameters, which are responsible for depth, width and resolution scaling respectively. These parameters are uniformly expressed by compound coefficient ϕ and constants α, β, γ which determinate search grids of parameters. Furthermore, as the FLOPS (floating point operations per second) of a regular convolution is proportional to $d \cdot w^2 \cdot r^2$, the constants
$alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$ have been chosen in the manner, that FLOP increment will be approximately proportional to 2^{ϕ}.

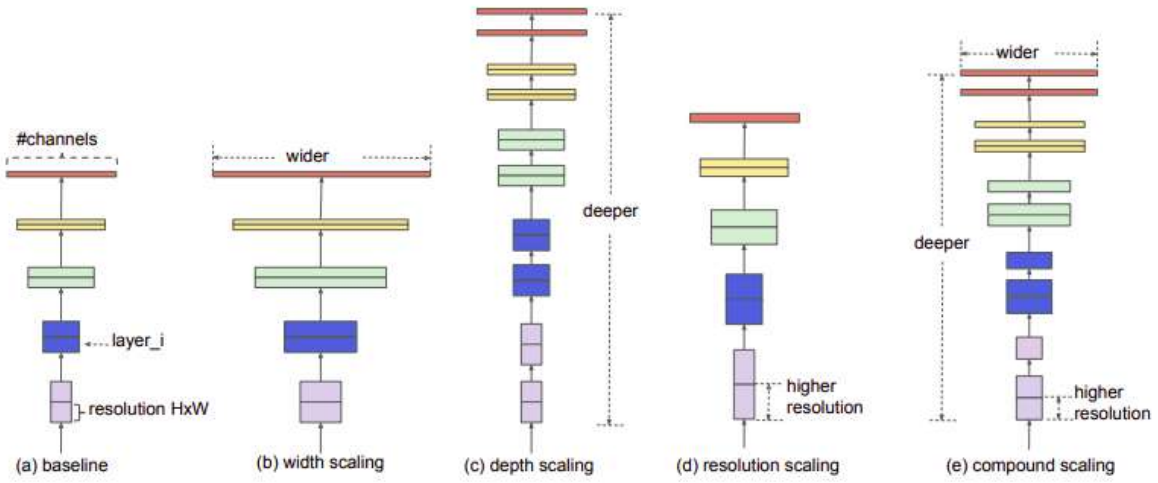


Figure 3.1: Model (a) scaling by width (b), depth (c), resolution (d), all dimensions (e), graphic source: [17].

$$
\begin{aligned}
depth : d &= \alpha^{\phi} \\
width : w &= \beta^{\phi} \\
resolution : r &= \gamma^{\phi} \\
s.t.\ \alpha \cdot \beta^2 \cdot \gamma^2 &\approx 2, \alpha \geqslant 1, \beta \geqslant 1, , \gamma \geqslant 1
\end{aligned} \tag{3.1}
$$

Where α, β, γ are already determined and constant for baseline model EfficientNet-B0, hence user is interested only in tuning ϕ, which should be set accordingly to available resources.

EfficientNet Architecture

Good model scaling method is an important part of the high performance CNNs models development, but to use its full potential, a high quality baseline network is required. In work [17], authors decided to develop an easily scalable mobile-sized network. In this purpose, multi-objective neural architecture search has been used with both accuracy and FLOP optimization task. Eventually, EfficientNet-B0 has been created, with the architecture presented in the table 3.1.

Table 3.1: EfficientNet-B0 baseline network architecture [17].

Stage	Operator	Resolution	Channels	Layers
1	Conv3x3	224 × 224	32	1
2	MBConv1, k3x3	112 × 112	16	1
3	MBConv6, k3x3	112 × 112	24	2
4	MBConv6, k5x5	56 × 56	40	2
5	MBConv6, k3x3	28 × 28	80	3
6	MBConv6, k5x5	14 × 14	112	3
7	MBConv6, k5x5	14 × 14	192	4
8	MBConv6, k3x3	7 × 7	320	1
9	Conv1x1 & Pooling & FC	7 × 7	1280	1

Each row in table 3.1 presents stage i, with its operator, layer's amount, input resolution and number of outputs channels. In the operator column there appears MBConv. MBConv (Mobile inverted bottleneck bloc [13]) denotes the block of multiple convolution, pooling, batch normalization and activation operations. In particular, MBConv widens tensor with first convolution using $1x1$ filter, then applies batch normalization

and activation function ReLU6, next on the obtained tensor uses depth-wise convolution with given kernel size (in this case $3x3$ or $5x5$), then once again applies batch normalization, followed by ReLU6 activation, and finally squeezes back the network with $1x1$ filter convolution to previously determined channel size and normalizes the batch for the last time. Additionally, there is skip connection, in order to give network the possibility to access previous activation without MBConv block modifications (prevents gradient vanishing in deep models). The amount of channel to, which the first block's convolution widens a tensor, is calculated with given input/output number of channels and expansion rate (in this case, 1 for stage 2, and 6 for the rests). The $narrow- > wide- > narrow$ architecture is presented on figure 3.2.

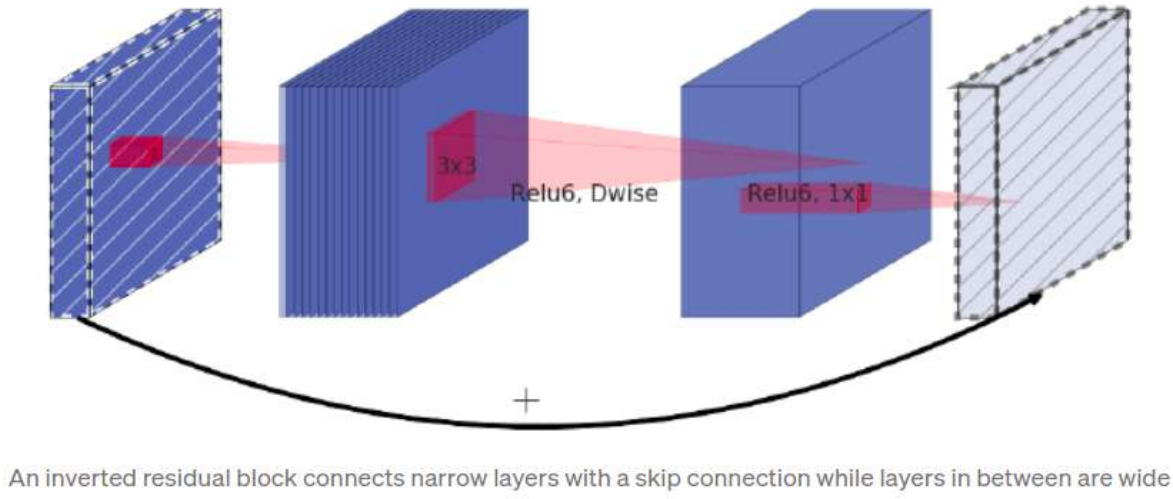


Figure 3.2: Mobile inverted bottleneck block MBConv, graphic source: [13].

EfficientNet-B2

With the high quality baseline EfficientNet-B0 3.1.2 and efficient scaling method 3.1.2, it was possible to create whole models EfficientNet family, what allows users to choose appropriate size of CNN to their available resources and problem difficulty. In the paper [17], there have been presented eight (including mobile-sized baseline network) models, with wide ranges of parameters, FLOPS and performance level. Table 3.2 is showing detailed information about these nets.

Table 3.2: EfficientNet models comparison [17], accuracy measured on ImageNet benchmark.

Name	Params	FLOPS	Top-1 Acc.
Efficientnet-b0	5.3M	0.39B	76.3
Efficientnet-b1	7.8M	0.70B	78.8
Efficientnet-b2	9.2M	1.0B	79.8
Efficientnet-b3	12M	1.8B	81.1
Efficientnet-b4	19M	4.2B	82.6
Efficientnet-b5	30M	9.9B	83.3
Efficientnet-b6	43M	19B	84.0
Efficientnet-b7	66M	37B	84.4

To visualize the gap between EfficientNet and state of the arts from that time, on figures 3.3, 3.4 there are presented plots of ImageNet accuracy versus amount of parameters and FLOPS, respectively.

The efficientNet's advantage in accuracy vs memory/computations over others popular models is clear, hence there was taken the decision about usage of this CNN's type in the problem discussed in this paper.

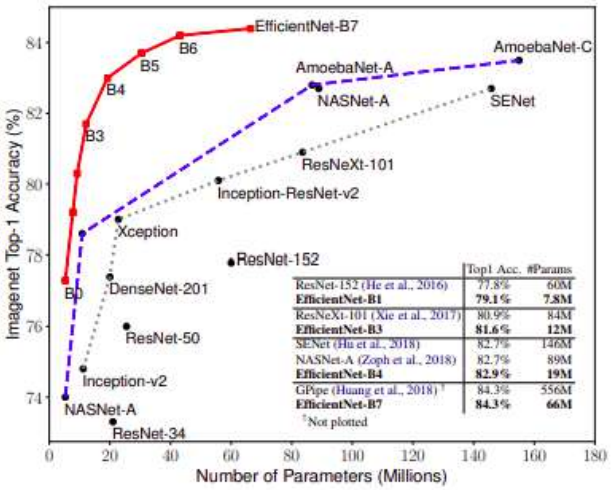


Figure 3.3: FLOPS vs. ImageNet accuracy, graphic source: [17].

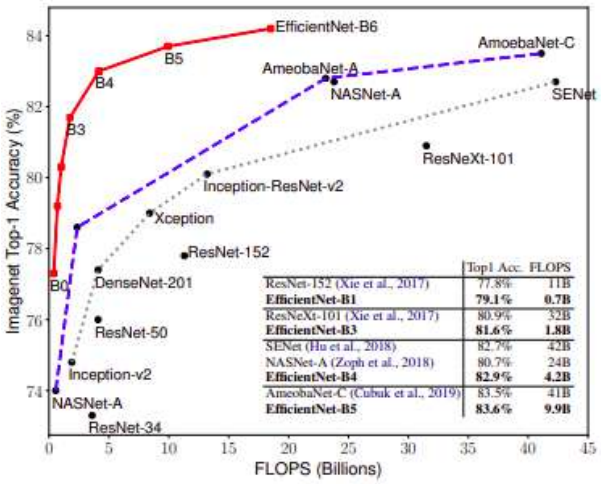


Figure 3.4: Model size vs. ImageNet accuracy, graphic source: [17].

In particular, version EfficientNet-B2 will be used to rulers recognition task on the coins from the considered dataset. Taking into account limited computational resources and the fact, that the dataset is rather small (1956 coins) in the context of deep learning training (even if the model will be already pretrained on ImageNet), the network can not be huge - otherwise underfitting would be serious threat. Hence, bigger models i.e. b4, b5, b6, b7 are not under the consideration. Table 3.3 shows direct comparison of remaining CNNs versions.

Table 3.3: EfficientNet b0, b1, b2, b3 models comparison, accuracy measured on ImageNet benchmark.

Name	Params	gain to b(i-1)	FLOPS	gain to b(i-1)	Top-1 Acc.	gain to b(i-1)
Efficientnet-b0	5.3M	-	0.39B	-	76.3	-
Efficientnet-b1	7.8M	+2.5M	0.70B	+0.31B	78.8	+2.5
Efficientnet-b2	9.2M	+1.4M	1.0B	+0.3B	79.8	+1
Efficientnet-b3	12M	+2.8M	1.8B	+0.8B	81.1	+1.3

As the baseline version b0 may be too small to obtain desirable results, it is reasonable to take the model which offers accuracy gain, not raising memory and computational cost significantly. In fact, version b1 offers such good accuracy-cost trade-off, however b2 also does not require much more FLOPS and parameters, thus, considering that b3 demands 80% more FLOPS than b2, Efficientnet-b2 has been chosen as the basic model for rulers recognition conducted in this paper. In order to check the correctness of the selection, models Efficientnet-b2 and Efficientnet-b0 have been compared in simple obverse's king task recognition (on dataset split training 85 %, test 15 %) and Efficientnet-b2 (82% Accuracy) outperformed Efficientnet-b0 (74% Accuracy) by 8% points on the test set, what confirms Efficientnet-b2 bigger capability to distinguish the coin's obverses.

3.1.3 Combined classification

The main problem to solve raised in this paper is, similarly to the previous work in this area [14], to recognize head/bust minted on obverse side of coins. However, this work's dataset, besides the Ancient Roman coins which were the subject in [14], contains specimens from 4 others medieval/early modern age countries. Furthermore, inspired by hierarchical approach from [8] which was used in reverse shape classification, in order to enhance main task performance, the country recognition based on reverse side will be conducted. These tasks are realized with modern CNN Efficientnet-b2 [17], which will be compared with results obtained by local feature method SIFT supported with the Bag of The Words technique and circle tiling presented in [2].

Straightforward classification

Before implementing more complicated ideas, it is always good to check out the most simple and direct ones, especially in deep learning where often simplicity is equal to good results. The most intuitive idea to recognize ruler present on obverse, is just putting obverse in CNN and choose the most probable option returned by softmax function (figure 3.5). The other straightforward idea is to pass both concatenated sides of the coin and hope that network will extract some helpful features from the reversed side image (figure 3.6).

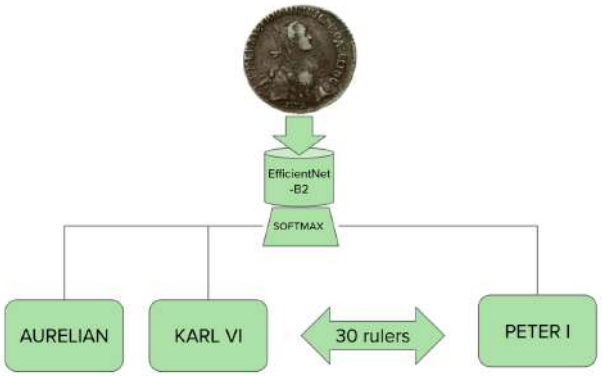


Figure 3.5: Ruler recognition using only obverse side.

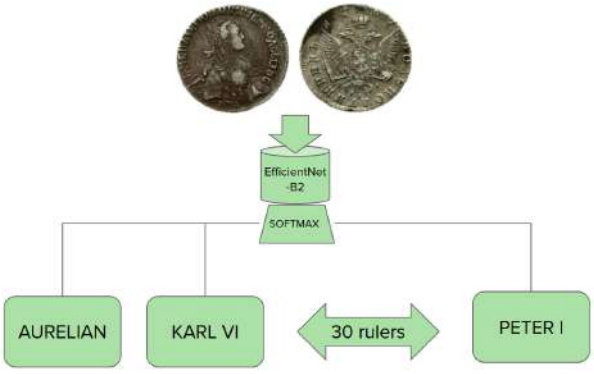


Figure 3.6: Ruler recognition using both coin sides.

However, both approaches are having some drawbacks, mostly resulting from small dataset availability. Overall, there are 30 classes distributed in 5 country super-classes to recognize. Additionally, all objects are the coins, what makes the classification task harder, as every class is distinguishable only by details (fine-grained problem). As dataset size is limited, there is serious risk of underfitting, even despite that smaller version of EfficientNet has been chosen and the net is already pretrained on huge ImageNet dataset. Especially, approach with passing two sides of coin to CNN can result on adding noise instead of additional feature, since it may be too hard for the model to learn anything useful. However, it is obvious, that reverse side indeed has some potential information, which could efficiently support obverse classification and it would not be reasonable to ignore that, by passing only side with ruler's bust.

Models combination

The solution, for proper incorporation of the reversed side to bust classification, is the usage of domain knowledge. Multiple previous works [3, 2, 4] have emphasized the importance of coin's context understanding by models, mostly in terms of reversed side. In paper [8] the domain knowledge, about which reverse's shape appears with which obverse's emperor, has been successfully incorporated, creating hierarchical model for reversed shape recognition, which notably outperformed straightforward reverse recognition. Inspired by this idea, this work presents combined classification.

The idea relays on two distinct EfficientNet-b2 models. One of them is simply, straightforward ruler classification, with usage of only obverse side. Second one is responsible for new sub-task - country recognition with usage of only reversed side. In the result, for the classified coin, there are obtained two softmax probabilities - one for ruler recognition (obverse), and second for country recognition (reverse). These probabilities are combined in the final score vector for each 30 rulers, where the biggest value is returned as a prediction. The scheme is visualized in figure 3.7.

Using reverse side in country recognition is kind of domain knowledge application, as this results from incorporating ruler-country relation and observation, that each country has fairly characteristic style and context of reverse (exemplary samples on figure 2.10):

- Russia - eagle of arms, legends in Cyrillic alphabet
- Austria - eagle of arms, characteristic shields
- Germany - wide spectrum of animals, persons, scenes and texts

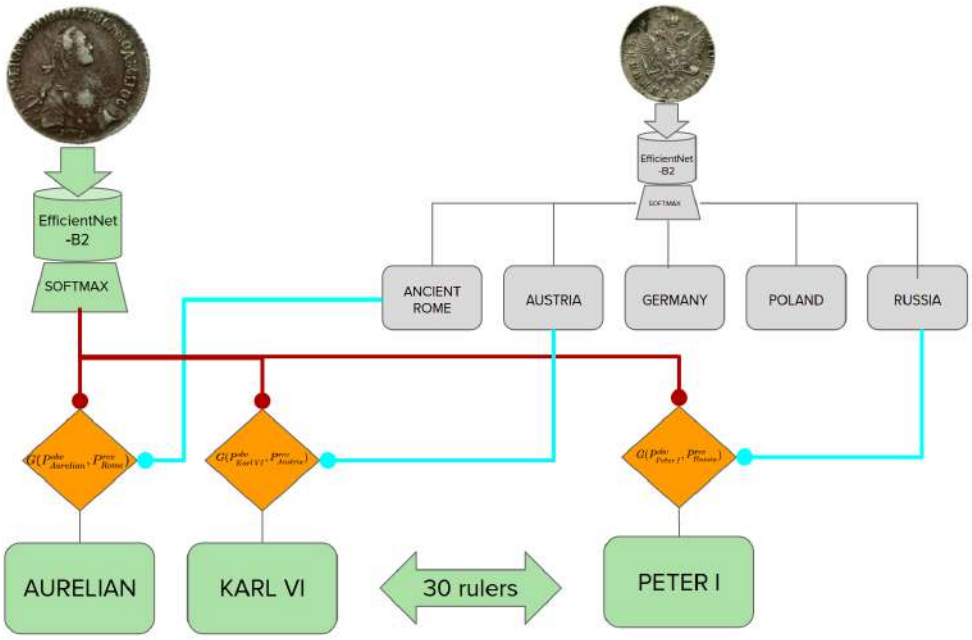


Figure 3.7: Combined model for ruler recognition.

- Poland - different types of shields and minor symbols
- Roman Empire - wide spectrum of animals, persons, scenes and texts

That knowledge gives grounds to believe, that country recognition should be highly efficient, and thus it will be possible to use the country information for increasing appropriate softmax values, obtained from obverse classification, what is expected to enhance overall performance. The method of combining two softmax vectors, in order to obtain final probabilities, is very intuitive and presented in definition 3.1.

Definition 3.1 (Softmax combination) *Let define the functions $P^{obv_j}(ruler_i)$, $P^{rev_j}(country_i)$, which denote the softmax values for given j-th coin's obverse obv_j and reverse rev_j sides, calculated for i-th ruler class $ruler_i$ and country class $country_i$. Then function G returns final score for $ruler_i$ and j-th coin:*

$$G(P^{obv_j}(ruler_i),P^{rev_j}(Ct(ruler_i))) = P^{obv_j}(ruler_i) \cdot I(P^{rev_j}(Ct(ruler_i)) \cdot c)) \tag{3.2}$$

where $Ct(ruler_i)$ returns origin country for $ruler_i$, c is a multiplier constant and I is given by $I(x) = 1$ for $x < 1$, $I(x) = x$ for $x \geqslant 1$. The function G is calculated 30 times (each ruler) for classified coin, thus score vector is created. The highest value in the vector is equivalent to ruler predicted to be present on coin's obverse.

$$return: \ MAX_{ruler_i}((G(P^{obv_j}(ruler_1),P^{rev_j}(Ct(ruler_1))), \ldots, G(P^{obv_j}(ruler_{30}),P^{rev_j}(Ct(ruler_{30})))))$$

Although the idea of combination of two tasks softmax is inspired by [8] hierarchical model, the way of combination is entirely different. Firstly, in work [8], the main task is reverse's shape recognition, and the combination relays on choosing only subspace of possible shapes indicated by previous obverse classification (the rest of shapes, which do not fit with predicted obverse's emperor, are rejected). This approach can propagate the obverse's recognition error to the final decision. In softmax combination method, proposed in this work, every reverse counts during final classification, thanks to function I, which does not allow multiplying $P^{obv_j}(ruler_i)$ with multiplier smaller than 1 - otherwise some values obtained from $P^{rev_j}(Ct(ruler_i)$ would be so close to 0, that $ruler_i$ would not be taken under consideration as the prediction, even if country would be predicted wrongly. Moreover, the influence of country model can be tuned by multiplier constant c, experiments (table 4.1) have shown, that $c = 1000$ is good choice, as it boosts chances of rulers from probable country, but does not dominate the rest of options in case if country has been predicted wrongly.

3.2 SIFT with Bag of Visual Words

In this section, there is briefly described SIFT+BoVW+SVM with circular tiling method [2], which has been used as a baseline approach during experiments performed in this work.

Scale Invariant Feature Transform (SIFT), from paper [10], relays on extraction of keypoints, from which further images descriptors are computed. These descriptors can be understood as distinctive, scale-invariant features, which allows reliable matching between different objects - in the case of this work, between coins. Below there are enumerated the major stages of image's features generation:

1. Scale-space extrema detection - algorithm searches over all scales and image locations, identifying candidates for scale and orientation invariant keypoints.
2. Keypoint localization - for each candidate, their location and scale are determined. Next, the keypoints are selected from the candidates, which seem to be the most stable according to some measures.
3. Orientation assignment - basing on local gradient's directions obtained from an image, one or more orientations are assigned to each keypoint.
4. Keypoint descriptor - the local image gradients, around each key-point, are measured and transformed into a key-point descriptor - feature representation, which is invariant for illumination changes and small shape distortions.

After feature generation, for every image, there is calculated histogram of SIFT features, what is achieved by applying Bag of Visual Words method. Bag of Visual Words (BoVWs) is the concept taken from text documents, where it is used to visualise the distribution of words present in the text. It has wide application in images processing, where an image can be represented using "visual" words. In the coin recognition problem, BoVWs is used to create histograms from SIFT keypoints descriptors. However, SIFT descriptors are having too big dimensionality to use them as the visual words, hence quantization method k-means is applied to reduce this dimensionality. In this work, there are created 50 clusters (visual words) from SIFT keypoints descriptors, as this amount is found to be the optimal one, basing on the experiments from paper [2].

SIFT features, although are very good at finding useful local keypoints in images, are lacking ability to extract image's spatial information. In order to, at least partly, overcome this problem for coin recognition, authors of [2] propose circular tiling technique. According to this idea, the image is divided on circular fragments, and then for each fragment there is calculated a distinct BoVWs histogram of SIFT descriptors, instead of calculating one histogram for the whole image. Thanks to that, each fragment's histogram can be interpreted as a some kind of spatial information, because it describes the overall composition of SIFT keypoints for given fragment. Moreover, circular tiling, in contrast to, for example rectangular tiling, is rotation invariant (because the coin is circular, so the fragments obtained from circular tiling do not change under rotation). On figure 3.8 there are presented three coin's fragments, obtained from circular tiling, with visible keypoints calculated by SIFT algorithm.

In this work, circular tiling is always used to obtain 3 fragments, as the value 3 is claimed to be optimal according to the authors of paper [2]. Experiments on the dataset discussed in this work have shown, that circular tiling increases SIFT+BoVW method accuracy, on country and ruler recognition task, by roughly 3% points in compare to SIFT+BoVW without circular tiling.

The whole SIFT+BoVWs method can be summered in the following steps, which are also visually presented on the figure 3.9:

1. Extract keypoints descriptors of SIFT local features from images in the dataset.
2. Use k-means technique in order to reduce dimensionality of descriptors set, obtaining as the result dictionary of K visual words.
3. Perform circular tiling in order to obtain fragmented images.
4. Calculate keypoints descriptors for every circular fragment and map them to appropriate clusters, accordingly to previously created dictionary of visual words.

5. Use the BoVWs technique and represent the images as concatenated histograms of visual words, taken from circular fragments. Each image should be represented with 3 concatenated histograms of visual words, as images are divided for 3 fragments by circular tiling method.

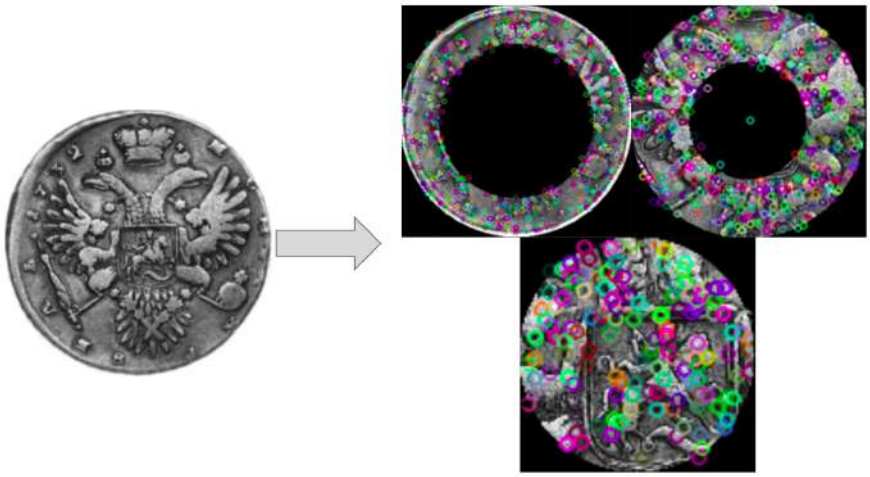


Figure 3.8: SIFT algorithm features extraction on coin's fragments obtained by usage of circular tiling.

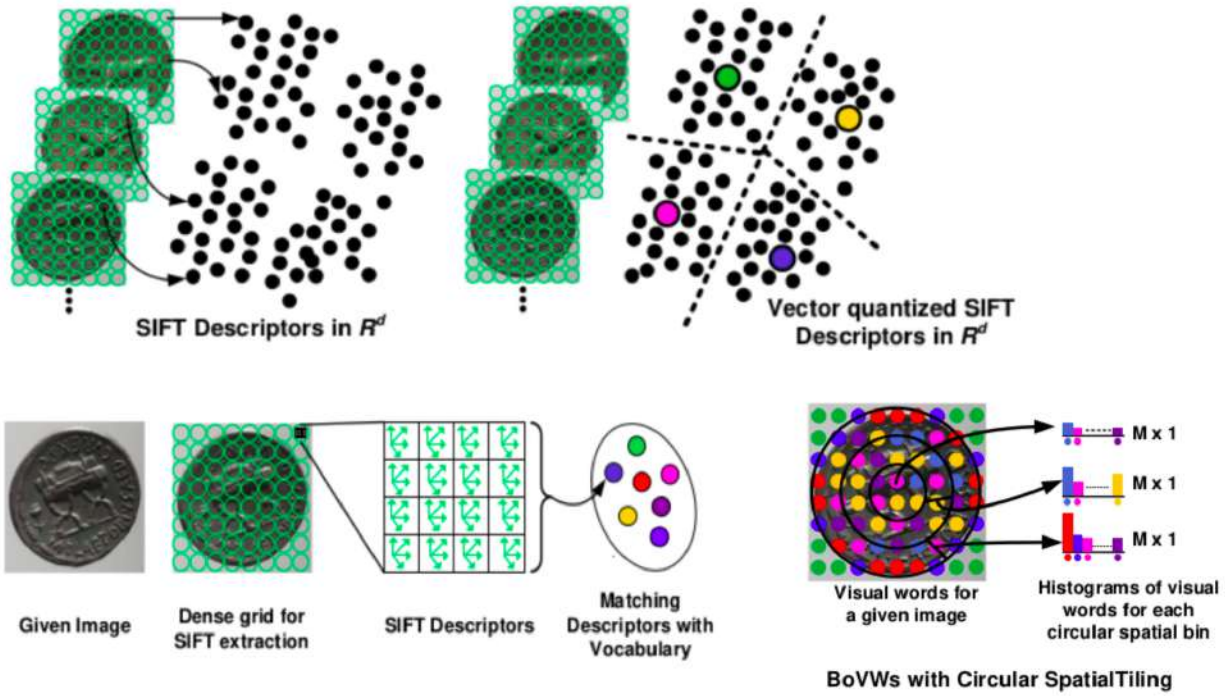


Figure 3.9: Visual explanation of baseline method SIFT+BoVW with circular tiling, graphic source: [2].

With the feature's histograms generated for every image, the last thing that is needed is a classifier. As the reliable classification method is crucial for the whole architecture's performance, within this work there are considered two different classifiers - support-vector machine (SVM) and k-nearest neighbors (KNN) algorithm. The first one, SVM, is based on finding hyperplanes, which divide a dataset into desired classes

in the best way. In that purpose, SVM changes the dimension of the data with usage of kernels, in particular Radial Basis Function (RBF) is used in this work (offered the best accuracy during experiments). The second classifier, kNN, is a very simple, but effective algorithm, which relays on the assumption, that similar objects are represented by the features in proximity according to a given measure. During experiments, SVM outperformed the kNN method by 8% points, thus it has been chosen to the final baseline architecture SIFT + BoVW + SVM.

4. Experiments

In this chapter there are presented and analyzed results of conducted experiments. Beforehand, there are described environment, methodology and motivation of researches. Furthermore, there is explanation of parameters choice and detailed interpretation of results in context of the problems discussed in this paper.

4.1 The aim and methodology of the experiments

In this section there is defined aim of the experiments. Afterwards, there are explanation of the training's methodology, way of model's evaluation and finally description of used technology and environment, in which experiments have been carried out.

4.1.1 Aim of the experiments

The experiments, which are presented in this part, have been conducted in order to answer the problems touched in this work. Primary goal is to check the performance of the modern CNN EfficientNet-b2 in ruler's bust recognition task conducted on the diverse dataset collected from Polish numismatic auctions, which contains coins, in wide spectrum of age and physical condition, with 30 different king's busts on obverse, belonging to 5 countries. In this purpose, EfficientNet-B2 will be used in three different methods:

- EfficentNet-b2 (obverse) - the model, which obtains only obverse coin side image on the input.
- EfficentNet-b2 (obverse‖reverse) - the model, which obtains image with both concatenated side.
- Combined EfficentNet-b2 - the model described in the section 3.1.3, which combines two CNNs and introduces new supporting task - coin's origin country recognition carried out on reverse side.

Performance of mentioned CNNs models will be compared with approach from paper [2] - local features SIFT extraction with bag of the visual words technique and circular tiling (SIFT + BoVW). In particular, the method Combined EfficentNet-b2 is expected to outperform other approaches, benefiting from domain knowledge incorporation - the links between kings busts and their origin country. Furthermore, the CNN's based methods capability to semantic coin understanding will be visualised with Grad-CAM technique. The country recognition task itself is very interesting in scientific way, as it will answer for the question if coins from different, although historically very close, countries like Austria, Germany, are easily distinguishable by deep learning techniques. Moreover, experiments will answer for the others research questions addressed in this thesis:

- Will deep learning based methods (EfficentNet-b2) outperform local feature based SIFT+BoVW method for considered dataset?
- Will domain knowledge incorporation, i.e. performing country recognition on reverse coin's side, enhance, or make worse the CNN's final obverse's bust/head recognition performance?
- Will deep learning based methods (EfficentNet-b2) be capable to understand semantic context of the coin and recognize obverse's bust/head with usage of elements, which would be in the field of interest for manual classification performed by human?

4.1.2 Methodology of the experiments

5-fold Cross Validation

The training has been carried out with usage of stratified 5-fold Cross Validation. Specifically, the entire dataset has been split in 5 folds, with around 390 coin's images each, in the way, that in the each part there is approximately equal number of coin's specimens with each king (possible 1 image difference per class, if total number of coins, with given ruler, is not dividable by 5). The fold split has been carried out by random sampling. This approach makes the results more reliable, as the model is tested with 5 different validations set (each fold becomes test set once, while the other folds are in the training set - in total 5 setups), instead of one, as it is the case in the traditional test-train data divide. Evaluation scores are always given as the averaged scores from every testing combination delivered by cross validation, and additionally, in some cases there are presented the detailed results for each test-fold.

Evaluation measures

The evaluation measures, which were used to evaluate models, are based on confusion matrix, what is the standard for the multi-label classification problems. In particular, there are used:

- Accuracy: $\text{acc} = \frac{1}{N}\sum_{k=1}^{|G|}\sum_{x:g(x)=k} I\left(g(x) = \hat{g}(x)\right)$, where N is the total number of observations, G is the set of the classes, I is the indicator, which returns 1 if the classes match and 0 otherwise.

- Macro averaged precision: $P_{\text{macro}} = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{TP_i}{TP_i+FP_i} = \frac{\sum_{i=1}^{|G|} P_i}{|G|}$, where G is the set of the classes, TP_i is "True Positive" for class i, FP_i is "False Positive" for class i.

- Macro averaged recall: $R_{\text{macro}} = \frac{1}{|G|}\sum_{i=1}^{|G|}\frac{TP_i}{TP_i+FN_i} = \frac{\sum_{i=1}^{|G|} R_i}{|G|}$, where G is the set of the classes, TP_i is "True Positive" for class i, FN_i is "False Negative" for class i.

- Macro averaged F1 score: $F1_{\text{macro}} = 2\frac{P_{\text{macro}} \cdot R_{\text{macro}}}{P_{\text{macro}}+R_{\text{macro}}}$.

Marco averaging has been chosen over the micro averaging, since it is essential in this task to obtain good performance for each classes, especially for these which are not highly represented, because in the case of collectables coins, rare coins usually are more valuable than common ones, hence it is important to recognize them well.

Technology and testing environment

The models have been trained with the usage of CUDA on the Google Colab GPU accelerator. All training and experiments have been carried out with $SEED = 1000$, in order to ensure researches reproducibility despite of randomness of certain processes. All programs within this work have been written in Python programming language, where particularly OpenCV's SIFT, Scikit-learn's evaluation measures and SVM, PyTorch's EfficientNet-b2 and data augmentation, Panda's data processing implementations have been used. The works were performed on Linux distribution's Ubuntu 20.04 operation system.

4.2 Experiments results and analysis

At the beginning of this section there are descriptions of data preprocessing and used augmentations. Next, there is explanation of parameters choice and afterwards there are presented results of conducted experiments, with proper analyze and statistical significance test. At the end of the section, there is summary of performed researches.

Preprocessing and data augmentation

The data do not need a lot of preprocessing, what results from the fact that images collected from auctions are already uniformly presented on the white background, with the obverse side on the left and concatenated reverse side on the right. Thus, the only thing to do at the beginning is cutting images in the middle and passing obverse and reverse sides of coin to respective models (beyond the unique model, which takes the concatenated images with both sides).

In the case of SIFT + BoVW models, the images are converted to greyscale, as it is SIFT requirement, and then resized to dimension (330,330), as this is the resolution of the smallest image in the dataset and for once it is big enough to maintain details and fits in the edge of available memory limit. Finally, there is performed circular tiling, in order to introduce spatial information to the SIFT models [2].

For the deep learning models, images are resized to the dimension (260,260), as the network EfficientNet is initialized with the parameters pretrained on ImageNet, where training data were passed in right that resolution - hence, in purpose to fully benefit from that initialization, it is good idea to maintain that shape, especially, when coin is circular shaped, so ratio $1x1$ is kept (once again, beyond the unique model with both sides at input, where ratio changes from $1x2$ to $1x1$). Beyond resizing, the data are normalized, as it is good practice in deep learning. From the same reason, as in image resizing, the means and STDs used for images normalization are taken from ImageNet training, thought it can be arguable if it is not better to calculate mean and STDs for actually used dataset. In order to clarify that uncertainty, experimentally has been proven that this choice has no significant influence on final accuracy in this particular case, but as it could be expected, model, where normalization was consistent with ImageNet pretraining parameters, was learning faster (because model's weights fit to normalization from very beginning of training), hence these means and STDs are the choice for further experiments. Finally, there are two augmentations, which are used only for the training data set (but not on the test), in order to increase training generalization. First one is the random horizontal flip with probability 50 %. Some coin types, within single classes, are having the bust once facing to the right and once facing to the left, what may be misleading for the model and thus learning such feature is undesired. Horizontal flip randomizes that physical aspect and makes model ignore it. The second augmentation is random image rotation in range of $[-5,5]$ degrees. Despite, that models assume uniform and clean image input, there are still present small rotations, as the coins are not positioned ideally. In order to make the model insensitive to such input noise, the training data are randomly rotated by a small degree. On a final note in the topic of transformations, the image is passed in the RGB format, although conversion to greyscale is very common in the CNN object classification. This is caused by the fact, that metal, which was used to mint the coin, thus associated with it colour, is often characteristic for certain regions and ages, so in order to not lose that information, the image in RGB is passed to the net. Similarly, the coin is not center-cropped to catch a particular bust region, as the legend on the edges, although often erased, is a very individual element for particular classes, so it could be costly to resign from such information.

Parameters choice for deep learning methods

An important part of successful training is choice of the parameters. There are not a lot of parameters in EfficientNet-b2 to set. The only thing that has been done in the model, was the change of the last classification layer to appropriate amount of outputs - for country recognition to 5 and for ruler recognition to 30, respectively to the amount of classes in both tasks. The other important side of the deep learning modelling is proper training's parameters setup. The dataset has been split on the batches sized 32 each. This value has been chosen, because this batch size, for a given image resolution (260x260), is close to the maximum memory, which fits into Google's GPU RAM limitations. In general big batch size was desired, because it ensures less frequent, but more robust parameters updates, and thus it encourages model to learn more generalized features, like obverse's bust or reverse's shape contexts, what is more desirable, in problem discussed in this work, than the small, detailed, differences, which would not by helpful in real-word numismatic environment. Generally, the batch size 32 is a fairly popular choice in the CNN training and the practice shows, that learning rate 0.01 pairs up with this batch size well. The reason behind this, is that learning rate 0.01 is relatively big, what makes a model to alter parameters significantly according to each update - what is a good idea, since batch size is big and should deliver not frequent, but good quality updates, thus it is reasonable to trust in that changes. There have been conducted experiments with others learning rate and batch size combinations, but the practice have shown that batch 32 and LR 0.01 is the best

choice, thus these values have been used in experiments. Furthermore, Adam optimizer and Cross Entropy Loss have been used, as these are acknowledged optimizer and loss function in deep learning environment for multi-label classification task.

Additionally, for the combined model discussed in the section 3.1.3 and its softmax combination method, there is a multiplier factor in formula 3.2, which is responsible for degree of country recognition's influence on final ruler's recognition softmax. In table 4.1, there are presented averaged classification accuracies, obtained from 5-cross-validation test-folds, depending from multiplier constant value. Regarding the results, multiplier constant $c = 1000$ ensures the best performance, hence this parameter's value is used in further tests.

Table 4.1: Multiplier constant parameter experiments for the softmax combination method.

Multiplier constant	1	10	100	1000	10000	100000
Avg. accuracy [%]	85.16	86.34	87.06	87.26	87.21	87.11

Parameters choice for SIFT+BoVW methods

In the second method's family, there is an entirely different set of parameters to choose because of the different algorithm's nature. For the first stage, local features extraction carried out by SIFT algorithm, there is not a lot to set, and all parameters have been left in default mode. In the second stage, features clustering in BoVW method is done by k-means clustering algorithm, where amount of cluster has been set to $k = 50$, accordingly to word's number recommended by authors of [2], where appropriate experiments have been carried out, in order to find out optimal k value. The last stage is classification back-end. For this role, two classifiers algorithm have been tried out - simply K-nearest neighbour (kNN) method and support vector machine (SVM), where SVM is taken with default scikit-learn framework parameters, in particular with radial basis function kernel, as it appeared to perform better, than linear kernel and sigmoid kernels, during initial experiments on the coin dataset from this work.

The table 4.2 presents comparison of SIFT+BoVW methods with SVM and kNN back-ends, carried out on country and ruler recognition tasks with cross validated dataset.

Table 4.2: Accuracy [%] comparison of SVM and kNN algorithms as the classification layers for SIFT + BoVW method.

Country recognition task						
Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg
SIFT + BoVW + SVM (reverse)	76.41	75.31	80.97	76.59	78.26	77.5
SIFT + BoVW + kNN (reverse)	67.69	64.63	72.23	70.99	72.63	69.63
Ruler recognition task						
Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg.
SIFT + BoVW + SVM (obverse)	48.97	47.07	50.12	50.89	49.87	49.38
SIFT + BoVW + kNN (obverse)	42.05	40.71	40.35	42.49	40.66	41.25

The results indicate clearly, that SVM outperforms the kNN algorithm in each task by approximately 8 % points, thus in the experiments SIFT + BoVW + SVM combination will be used.

4.2.1 Experiments results

The first table 4.3, presents the comparison of previously discussed models performances in ruler recognition task, and additionally the results of country recognition side-task, obtained by EfficientNet-b2, which are compared to SIFT + BoVW + SVM baseline technique. The models are evaluated with usage of accuracy metric, calculated for every test-fold from. considered in this work, coin dataset. Afterwards, there have been performed CNN Grad-CAM visualisations and result's statistical significance analyze. Finally, there is summary of experiments and conclusions.

The results (4.3) show, that proposed combined EfficientNet-b2 model outperforms the rest of models in every test-fold for ruler recognition task. In particular, the average accuracy of combined model is

Table 4.3: Models accuracy [%] for country and recognition tasks.

Country recognition task						
Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg.
SIFT + BoVW + SVM (reverse)	76.41	75.31	80.97	76.59	78.26	77.5
EfficentNet-b2 (reverse)	95.64	94.14	94.85	95.92	94.62	95.03
Ruler recognition task						
Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg.
SIFT + BoVW + SVM (obverse)	48.97	47.07	50.12	50.89	49.87	49.38
EfficentNet-b2 (obverse)	86.66	84.73	84.57	84.73	85.16	85.17
EfficentNet-b2 (obverse$\|\|reverse$)	80.76	77.6	81.49	86.0	85.16	82.2
Combined EfficientNet-b2 model	88.71	84.98	86.63	89.82	86.18	87.26

higher by 2.09 % points, than the second-best method - EfficientNet-b2 with obverse input. Moreover, the EfficientNet-b2, with concatenated obverse and reverse images input, is worse than 'only obverse' CNN model by average 2.97 % points and worse than combined model by 5.06 % points. The non-deep learning method SIFT+BoVW+SVM obtains much worse results than all CNN nets for both considered tasks.

Continuing on, there is table 4.4, which presents the specific metric measures values for two best methods - combined EfficientNet-b2 model and EfficientNet-b2 (ovberse input), without incorporated country-domain knowledge. The metric are averaged for every 5 test-fold and for each of them, combined model has better score in ruler recognition task, by approximately 2% points, than the other model.

Table 4.4: Comparison of obverse's EfficentNet-b2 model and combined EfficentNet-b2 model average performances.

Country recognition task				
Method	Precision	Recall	F1 score	Accuracy
EfficentNet-b2 (reverse)	95.07	94.89	94.94	95.03
Ruler recognition task				
Method	Precision	Recall	F1 score	Accuracy
EfficentNet-b2 (obverse)	87.01	84.47	84.58	85.17
Combined EfficientNet-b2 model	88.63	86.55	86.69	87.26

The table 4.5 contains the results for country recognition task. Country recognition is not the main purpose of this work, nevertheless it is responsible for enhancement of ruler recognition in the proposed combined EfficentNet-b2 approach, hence it is essential to obtain high performance in this sub-task. The EfficentNet-b2 (with reverse coin's side input), which is incorporated in the combined model, significantly outperforms the baseline SIFT + BoVW + SVM (with reverse coin's side input) in terms of average 5-cross-validation accuracy. It is worth to notice, that Roman Empire's coins are clearly better recognized, than the ones from other country. This fact confirms, that ancient Roman Empire's coins looks differently, due to age difference and thus another style, materials, coins condition and technique of minting, in compare to medieval coins. Moreover, for the medieval countries there is visible over 20% point improvement with usage of CNN's method. As Roman Empire is easily distinguishable by local features SIFT model, it is confused with similarly looking medieval coins. The CNN's advantage in understanding spatial information and actual motif context may be the reason of this big improvement.

Table 4.5: Accuracy [%] per class for EfficentNet-b2 and SIFT based models, country recognition task.

Method	Austria	Germany	Poland	Roman Empire	Russia
EfficentNet-b2 (reverse)	92.91	94.19	95.58	98.59	93.22
SIFT + BoVW + SVM (reverse)	70.37	72.79	74.89	92.98	74.68

Table 4.6 presents, in what degree particular classes have improved their intra-class recognition accuracy by using the combined model, in compare to EfficentNet-b2 (obverse input) and SIFT + BoVW + SVM

(obverse input) models. Regarding the SIFT based model and deep learning nets comparison, every class have noted big improvement. The situation is different in comparison between two CNN models. Most of the classes (19) have improved their intra-class recognition after applying the combined model, 10 classes did not change their accuracy, and one class, Maximilian III Jose, decreased its classification accuracy by 1.66% point. On the other hand, John III Sobieski is the class which has improved the most, by 10.33% points.

Table 4.6: Accuracy [%] per class for ruler recognition task. Models M0: SIFT + BoVW + SVM (obverse), M1: EfficentNet-b2 (obverse), M2: Combined country/ruler model.

Ruler	Country	M0	diff. to M2	M1	diff. to M2	M2
Alexander III	Russia	75.6	-22.85	98.46	0.0	98.46
Anna Ioannovna	Russia	31.42	-37.14	65.23	-3.33	68.57
Elizabeth	Russia	56.08	-37.41	92.16	-1.33	93.5
Nicholas II	Russia	79.24	-17.42	96.66	0.0	96.66
Catherine II	Russia	43.71	-48.2	91.92	0.0	91.92
Peter I	Russia	33.33	-57.77	88.88	-2.22	91.11
Augustus III the Sas	Poland	73.98	-22.61	95.42	-1.17	96.6
Sigismund I Old	Poland	0.0	-76.0	72.0	-4.0	76.0
Sigismund III	Poland	63.15	-28.42	89.47	-2.1	91.57
Stephen Bathory	Poland	65.76	-27.69	93.46	0.0	93.46
John II Casimir	Poland	36.83	-44.26	77.57	-3.52	81.1
John III Sobieski	Poland	49.33	-34.91	73.91	-10.33	84.25
Karl VI	Austria	58.75	-37.5	96.25	0.0	96.25
Leopold I	Austria	54.44	-36.66	88.88	-2.22	91.11
Leopold V	Austria	17.45	-70.72	84.36	-3.81	88.18
Ferdinand II	Austria	15.99	-47.99	58.0	-5.99	63.99
Ferdinand III	Austria	28.66	-53.33	80.0	-2.0	82.0
Franz Joseph I	Austria	52.0	-42.66	89.33	-5.33	94.66
Friedrich Wilhelm I	Germany	78.45	-17.86	93.97	-2.35	96.32
Friedrich Wilhelm II	Germany	49.25	-38.99	85.58	-2.66	88.25
Friedrich Wilhelm III	Germany	36.66	-49.99	83.33	-3.33	86.66
Ludwig I	Germany	55.55	-41.94	97.5	0.0	97.5
Ludwig II	Germany	31.66	-60.83	92.5	0.0	92.5
Maximilian III Jose	Germany	49.1	-40.89	91.66	1.66	90.0
Neron	Roman Empire	40.0	-30.76	67.69	-3.07	70.76
Philip I	Roman Empire	47.14	-25.71	72.85	0.0	72.85
Trajan	Roman Empire	33.41	-38.41	68.08	-3.74	71.83
Aurelian	Roman Empire	66.19	-17.42	83.61	0.0	83.61
Constantine the Great	Roman Empire	42.85	-43.14	84.66	-1.33	86.0
Hadrian	Roman Empire	26.81	-54.17	80.98	0.0	80.98

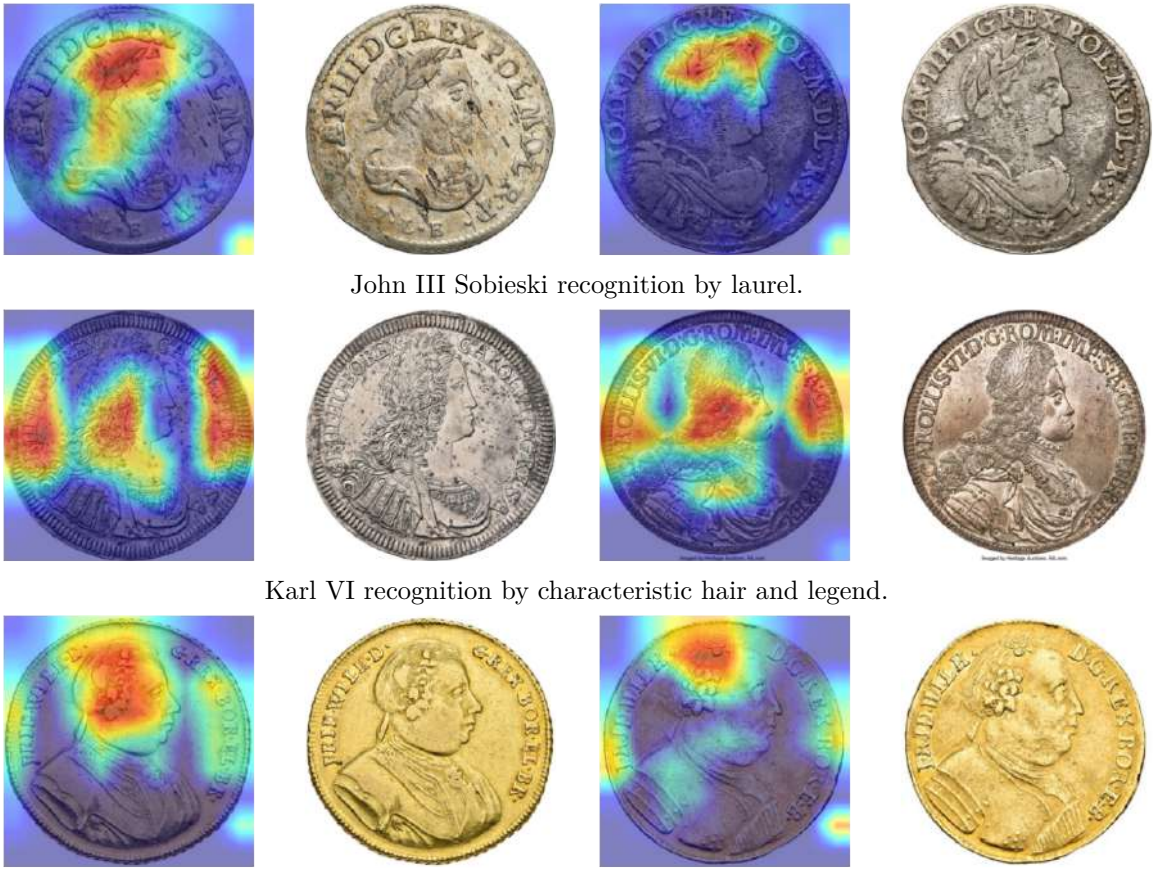
In order to better understand, why, in the proposed combined model and EfficentNet-b2 (obverse input) model comparison, John III Sobieski is recognized 10.33 % points better by the combined model and Maximilian III Jose denoted small decrease of accuracy, table 4.7 is presenting part of the confusion matrix, which shows false positives (FP) and false negatives (FN) for these two classes and methods. Particularly interesting are gray-highlighted cells, which indicate the confusion matrix's differences between the two models. For John III Sobieski, the FN/FP decrease can be observed only among non-polish classes, where the FN/FP gain is present only among polish rulers. The same pattern (but with German FN/FP gains, instead of Polish FN/FP gains) is observed for Maximilian III Jose, with one exception for Ferdinand II false negative, where amount increased by one (Maximilian III Jose is from Germany, so for others country value is expected to decrease - example of how the combined model's country miss-classification, between two similarly countries - Germany and Austria, propagates the error to ruler miss-classification). This phenomenon confirms, that country recognition incorporated by the combined model, in fact reduces potential ruler's search space to the kings from recognized country, what results with bigger amount of FN/FP among the predicted country's rulers,

but also significantly reduces FN/FP among the others countries and eventually increases overall accuracy of ruler recognition task. Maximilian III Jose is the only class, which noted decrease of intra-class accuracy (by 1.66 % points), but in fact, confusion matrix is showing, that total amount of false negatives (from which intra-class accuracy is calculated) for Maximilian III Jose class, has not changed after the combined model application (5 FN vs 5 FN). However, Ferdinand II's false negative replaced Aurelian's false negative, and as these two coins appear to belong to different folds, it has led to the difference in accuracy averaging due to small differences of class specimen's amounts among the folds.

Table 4.7: Confusion matrix for chosen rulers R1: John III Sobieski R2: Maximilian III Jose and models M1: EfficentNet-b2 (obverse) M2: Combined country/ruler model. Matrix is the sum of all five folds occurrences, rows with zeros values only have been deleted. Highlighted cells indicate differences between models.

False negatives for R1: John III Sobieski R2: Maximilian III Jose					
Ruler	Country	M1 & R1	M2 & R1	M1 & R2	M2 & R2
Catherine II	Russia	1	0	0	0
Sigismund III	Poland	1	1	1	1
Stephen Bathory	Poland	1	1	0	0
John II Casimir	Poland	3	4	0	0
John III Sobieski	Poland	-	-	0	0
Augustus III the Sas	Poland	0	2	0	0
Ferdinand II	Austria	0	0	0	1
Ferdinand III	Austria	1	0	0	0
Karl VI	Austria	1	0	0	0
Leopold V	Austria	0	0	1	1
Friedrich Wilhelm I	Germany	3	1	0	0
Friedrich Wilhelm II	Germany	3	0	2	2
Friedrich Wilhelm III	Germany	1	1	0	0
Maximilian III Jose	Germany	2	1	-	-
Philip I	Roman Empire	1	0	0	0
Aurelian	Roman Empire	0	0	1	0
Constantine the Great	Roman Empire	2	0	0	0
False positives for R1: John III Sobieski R2: Maximilian III Jose					
Ruler	Country	M1 & R1	M2 & R1	M1 & R2	M2 & R2
Peter I	Russia	0	0	2	0
Catherine II	Russia	0	0	1	0
Sigismund III	Poland	0	1	0	0
John III Sobieski	Poland	-	-	2	1
Ferdinand III	Austria	1	1	0	0
Friedrich Wilhelm II	Germany	1	1	2	2
Maximilian III Jose	Germany	0	0	-	-
Trajan	Roman Empire	1	0	1	1
Aurelian	Roman Empire	1	0	0	0

Finally, in purpose to check if deep learning methods are capable to recognize coins in similar way to human, CNN visualization technique Gradient-weighted Class Activation Mapping (Grad-CAM) [15] has been used. Grad-Cam method utilizes the gradients of the classification score, obtained from the final convolutional feature map. In that way, there are identified the parts of an input image, that most impact the classification score. The heatmaps on figure 4.1 are presenting visualizations performed for three kings recognition. These examples confirm, that CNN is able to find characteristic parts of the coin, like John III Sobieski's laurel or Karl VI's and Friedrich Wilhelm I's curly hair. Classifiers also are interested in legends, as it is a fairly individual element of every coin. Therefore, the visualizations are consistent with [4, 8] claim, that CNNs used in coin recognition task are capable to spot regions, which would be used during manual classification.



John III Sobieski recognition by laurel.

Karl VI recognition by characteristic hair and legend.

Friedrich Wilhelm I recognition by curls.

Figure 4.1: EfficientNet-b2(obverse) heatmap visualization delivered by Grad-CAM technique [15].

Statistical tests

In order to check the statistical significance of performed experiments, the Friedman test with Shaffer's post hoc has been carried out. The test has been conducted on the 5 test-fold's, using accuracy results for ruler recognition task, obtained by four methods presented in the table 4.3.

Every method's average rank (calculated from 5 folds distinct ranks) outputted by Friedman procedure, are presented in the table 4.8. Friedman statistic considering reduction performance (distributed according to chi-square with 3 degrees of freedom) is 13.74. Knowing that, it can be inferred, that p-vaue computed by Friedman Test is equal to $p = 0.00328$, what means, that it is statistically safe to assume, that the methods performances do not belong to the same distribution and thus are classifying the coins in different, sometimes better (higher ranks) or worse (lower ranks), manner.

Table 4.8: Average Rankings of the algorithms.

Algorithm	Ranking
Combined EfficentNet-b2 model	1
EfficentNet-b2 (obverse)	2.3
EfficentNet-b2 (obverse$\|\|reverse$)	2.7
SIFT + BoVW + SVM (obverse)	4

In order to obtain more specific information regarding each method, Shaffer's post hoc has been applied. Table 4.9 is presenting results achieved on post hoc comparisons for $\alpha = 0.05$, where α indicates an acceptable

probability, that difference in compared method's performance is not accidental and experiments are statistically significant. In particular, in each row of the table, there are Shaffer-values, which are the thresholds for p-values i.e. the difference between compared methods is statistically significant, if p-value $\leqslant$ Shaffer-value. Moreover, in the table, there are z values, which denotes the difference between an observed statistic and its hypothesized population parameter in units of the standard deviation. Results are showing, that only the difference between the combined model and SIFT+BoVW+SVM (obverse) is statistically significant for $\alpha = 0.05$.

Table 4.9: P-values Table for $\alpha = 0.05$

i	algorithms	$z = (R_0 - R_i)/SE$	p	Shaffer
6	SIFT + BoVW + SVM (obverse) vs. Combined model	3.674235	0.000239	0.008333
5	SIFT + BoVW + SVM (obverse) vs. EfficentNet-b2 (obverse)	2.082066	0.037336	0.016667
4	EfficentNet-b2 (obverse$\|reverse$)$vs.Combinedmodel$	2.082066	0.037336	0.016667
3	SIFT + BoVW + SVM (obverse) vs. EfficentNet-b2 (obverse$\|reverse$)	1.592168	0.111347	0.016667
2	EfficentNet-b2 (obverse) vs. Combined model	1.592168	0.111347	0.025
1	EfficentNet-b2 (obverse) vs. EfficentNet-b2 (obverse$\|reverse$)	0.489898	0.624206	0.05

This paper assumes, that only tests with $\alpha = 0.05$ statistical significance level are reliable. Under that criteria, only superiority of combined EfficientNet-b2 method over the local features based algorithm is proven. Regarding the rest of comparisons, conducted experiment can be considered as not sufficient, thus these methods need more tests in order to indicate their hierarchy.

4.2.2 Experiments summary

Purposes of the experiments, stated in the section 4.1, have been achieved. Similarly the research questions addressed for these tests have been answered. In particular, there have been shown in the experiments, that for the considered multi-country data set, deep learning net EfficientNet-b2, similarly like others CNNs in case of the Ancient Roman coins datasets in related works [1, 8], ensures much better recognition performance, than method based on local features SIFT. Also, the incorporation of domain knowledge about kings-country links and easiness of coin's origin country recognition by reverse coin's side (because country-characteristic reverse side elements), allowed the combined model to overcome the EfficentNet-b2 (obverse) model in king's recognition task. The performance of EfficentNet-b2 (obverse‖reverse) have shown, that adding reverse coin's side without any domain knowledge clue for the training (like country recognition sub-task for combined model), does not only not improve ruler's recognition, but made it worse in compare to only 'obverse input' CNN method. This is because, this raw reverse information has occurred to be too complicated for CNN model, without any additional obverse-association context, thus instead of enriching the classification with new, useful features, the reverse side became noise, which made obverse bust recognition harder and in result the EfficentNet-b2 (obverse‖reverse) model appeared to be underfitted. The digital analyze of the proposed combined method have shown, that incorporation of country recognition sub-task successfully limits the search space of potential rulers present on obverse, to the rulers from predicted country, what taking in the account high performance of country recognition, improves overall model's performance. Moreover, the CNN visualization, carried out with usage of Grad-CAM technique, showed that the model is capable to spot characteristic elements, which could be taken under consideration by human expert during manual recognition. Finally, Friedman test with Shaffer's post-hocs has been carried out on the results, in order to check statistical significance of performed researches. It occurred, that only the superiority of the combined model over SIFT+BoVW+SVM traditional approach has been proven with sufficient statistical significance level $\alpha = 0.05$ threshold (significance with $\alpha = 0.05$ level is assumed to be necessary in order to make method comparison reliable), what implies the need of further experiments on additional dataset, in order to fulfill statistical reliability requirement.

5. Conclusions and future work

This work's main aim was to develop modern, deep learning based method, in order to perform recognition of ruler's head/bust minted on the obverse side of the coins, belonging to the most common collectable coins types occurring in the Polish numismatic trade market. Furthermore, the method was supposed to benefit from incorporated domain knowledge and outperform baseline methods, i.e. SIFT+BoVW from paper [2] and simple deep learning methods without incorporated domain knowledge. This objective has been entirely fulfilled by the presented combined EfficientNet-b2 model. This method, uses acknowledged, very high quality CNN EfficientNet in version, carefully scaled to the problem and available resources, EfficientNet-b2. Moreover, the combined method incorporates domain knowledge about obverse's ruler bust and coin's origin country association, what was motivated by the observation, that each country has fairly characteristic reverse motifs and manufacturing style. Thus, the relatively easy sub-task of reverse's country recognition, successfully reduced the potential rulers search space for the final obverse recognition task, what resulted in average accuracy improvement by 2.09% points, in respect to the simple EfficentNet-b2 with coin's obverse input, by 5.06% points, in respect to the simple EfficentNet-b2 with concatenated coin's obverse|reverse input and finally by 37.88% points, in respect to the SIFT+BoVW+SVM local features method with coin's obverse input. These results allow responding to research questions addressed in this work. In particular, EfficentNet-b2 occurred to be appropriate deep learning method for obverse's bust/head recognition for collectable coins dataset and managed to highly outperform local feature based method. Furthermore, incorporating knowledge about coin's origin countries into the combined model architecture, allowed enhancing performance in comparison to the others methods. Moreover, what is highly important for the potential real-life application of the automatic coin recognition, the visualization of CNN method delivered by Grad-CAM technique showed, that model was capable to spot coin's characteristic elements, which could be consistent with the human's choice elements used for the coin recognition. The meaningful difference between researches conducted within this thesis and related works, is that the concerned dataset consist not only from the Ancient Roman coins, like it is the case in the mayor part of the related researches, but it has been enriched with coins from other ages and regions, making it more diverse. Specifically, for purposes of this work, the labeled, balanced dataset has been created, which includes nearly 2000 coins with good quality coin's obverse and reverse sides images, from 5 different countries and with 30 classes of rulers head/bust minted on the obverse side. Furthermore, this dataset has been sampled from Polish numismatic online auctions, hence is very useful in terms of conducted researches for practical applications for this environment.

The obtained results can be useful contribution for collectables coin recognition, especially in the context of the latest deep learning methods usage in the recognition task of various coin types from ancient, medieval and early modern ages. There are multiple ways of continuing and developing works carried out in this thesis. Firstly, taking into the consideration, that statistical Friedman test indicated, that obtained result for some method's comparison are not sufficiently statistically significant, it is reasonable to collect bigger datasets and perform more experiments, in order to ultimately confirm conclusions obtained in this thesis. Furthermore, as deep learning is developing rapidly and there are constantly appearing newer and better CNNs (at the moment of finishing this thesis, paper [18] have introduced EfficientNetV2 family), it is reasonable to think, that replacing the EfficientNet-b2 CNN with newer one in the proposed combined model, will result with performance improvement. Also, application of another transformations set during the training, for example random region cropping or grayscale conversion, may results, although not necessarily, in the better generalization and context understanding abilities. There are also different interesting directions in the collectable coin's area, like automatic evaluation of coin's condition, what could be potentially incorporated in the coin recognition architecture model and thus enhance classification performance, by taking coin's condition into the consideration.

Bibliography

[1] H. Anwar, S. Anwar, S. Zambanini, F. Porikli. Deep ancient roman republican coin classification via feature fusion and attention. *Pattern Recognition Volume 114, June 2021, 107871*, 2021.

[2] H. Anwar, S. Zambanini, M. Kampel. A bag of visual words approach for symbols-based coarse-grained ancient coin classification. *OAGM/AAPR 2013 proceedings*, 2013.

[3] H. Anwar, S. Zambanini, M. Kampel, K. Vondrovec. Ancient coin classification using reverse motif recognition. *IEEE SIGNAL PROCESSING MAGAZINE [64] july 2015*, 2015.

[4] J. Cooper, O. Arandjelović. Understanding ancient coin images. *Recent Advances in Big Data and Deep Learning (pp.330-340)*, 2019.

[5] P. Foret, A. Kleiner, H. Mobahi, B. Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *ICLR 2021*, 2021.

[6] K. He, X. Zhang, S. Ren, J. Sun. Deep residual learning for image recogni- tion. *CVPR, 2016, pp. 770–778*, 2016.

[7] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger. Densely connected convolutional networks. *CVPR, 2017, pp. 4700–4708*, 2017.

[8] J. Kim, V. Pavlovic. Discovering characteristic landmarks on ancient coins using convolutional networks. *International Conference on Pattern Recognition (ICPR)*, 2015.

[9] A. Krizhevsky, I. Sutskever, G. Hinton. Imagenet classification with deep convolu- tional neural networks. *Advances in Neural Information Processing Systems pp. 1097–1105 (2012)*, 2012.

[10] D. G. Lowen. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision, 60:91–110.*, 2004.

[11] H. Pham, Z. Dai, Q. Xie, M.-T. Luong, Q. V. Le. Meta pseudo labels. *Google AI, Brain Team, Mountain View, CA 94043*, 2021.

[12] M. Reisert, O. Ronneberger, H. Burkhardt. An efficient gradient based registration technique for coin recognition. *Proc. MUSCLE CIS Coin Competi- tion Workshop, 2006, pp. 19–31*, 2006.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510-4520*, 2018.

[14] I. Schlag, O. Arandjelović. Ancient roman coin recognition in the wild using deep learning based recognition of artistically depicted face profiles. *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Conference on Computer Vision (ICCV'17)*, 2010.

[16] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv, page 1409.1556.*, 2014.

[17] M. Tan, Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning, 2019*, 2019.

[18] M. Tan, Q. V. Le. Efficientnetv2: Smaller models and faster training. *International Conference on Machine Learning, 2021*, 2021.

[19] L. van der Maaten, P. Boon. Coin-o-matic: A fast system for reliable coin classification. *In Proc. MUSCLE CIS Coin Recognition Competition Workshop, pages 7–18*, 2006.

[20] M. Zaharieva, M. Kampel, , S. Zambanini. Image-based recognition of ancient coins. *Proc. Int. Conf. Computer Analysis of Images and Patterns (CAIP), 2007, pp. 547–554*, 2007.