Ancient Roman Coin Recognition in the Wild using Deep Learning Based Recognition of Artistically Depicted Face Profiles

Imanol Schlag and Ognjen Arandjelović†
School of Computer Science
University of St Andrews
Scotland, United Kingdom
†ognjen.arandjelovic@gmail.com

Abstract

As a particularly interesting application in the realm of cultural heritage on the one hand, and a technically challenging problem, computer vision based analysis of Roman Imperial coins has been attracting an increasing amount of research. In this paper we make several important contributions. Firstly, we address a key limitation of existing work which is largely characterized by the application of generic object recognition techniques and the lack of use of domain knowledge. In contrast, our work approaches coin recognition in much the same way as a human expert would: by identifying the emperor universally shown on the obverse. To this end we develop a deep convolutional network, carefully crafted for what is effectively a specific instance of profile face recognition. No less importantly, we also address a major methodological flaw of previous research which is, as we explain in detail, insufficiently systematic and rigorous, and mired with confounding factors. Lastly, we introduce three carefully collected and annotated data sets, and using these demonstrate the effectiveness of the proposed approach which is shown to exceed the performance of the state of the art by approximately an order of magnitude.

1. Introduction

The present is an exciting time for computer vision: the field itself has matured, the hardware needed to support developed algorithms is affordable and pervasive, and the potential user base is greater than ever owing to the increasing recognition of the benefits that machine intelligence can offer. This technological and social climate has opened a vast field of potential new applications for computer vision, with many attractive and exciting problems emerging from its applications in arts and humanities. In this work we are interested in the application of computer vision to ancient numismatics and the classification of Roman Imperial coins in particular.

1.1. Terminology

Considering the interdisciplinary nature of the present paper, it is important to explain the relevant numismatic terminology so that the specific task at hand and its challenges can be clearly understood. A succinct summary is presented next.

Firstly, when referring to a *coin*, the reference is made to a specific physical object i.e. a specimen. This is to be contrasted with a coin *type*. A coin type is a more abstract concept which is characterized by the semantic features shown on both sides of the coin (the obverse i.e. the "front", and the reverse i.e. the "back"). Multiple coins of the same type have the same visual elements e.g. the head or bust of a particular emperor with specific clothing (e.g. drapery or cuirass, crowned or laureate) and *legends* (textual inscriptions), a particular reverse motif, etc. Notice that although the visual elements on coins of the same type are semantically the same, their depictions may differ somewhat. The reason lies in the fact that the same coin type was minted using dies created by different engravers. For example, observe in Figure 2 which shows three specimens of the same type, that the spatial arrangements of the legend (by definition the same in all cases) is different between the very fine example in Figure 2(b) and the extra fine example in Figure 2(c). In the former case the break (space) in the legend is AEQVITA-SAVG, and in the latter AEQVI-TASAVG. Nevertheless the type is the same.

1.1.1 Condition grades

An important consideration in the analysis of ancient coins concerns their *condition*. Succinctly put the condition describes the degree of preservation of a coin, or equivalently the amount of damage it suffered since it was minted. The usual grading scale adopted in ancient numismatics includes the following main grades: (i) poor, (ii) fair, (iii)

good, (iv) very good, (v) fine, (vi) very fine, and (vii) extremely fine. Virtually universally (i.e. save for extremely rare coin types) only the last three are considered of interest to collectors, that is fine (F), very fine (VF), and extremely fine (EF or XF). Note that less frequently used transitional grades can be derived from the main seven by qualifiers e.g. near or almost fine (nF, aF), better than fine (F+), etc.

An ancient coin in a fine condition displays all the main visual elements of the type, as illustrated with an example in Figure 2(a). A very fine coin also has more subtle elements preserved such as clothing creases as exemplified in Figure 2(b). An extremely fine condition coin is in approximately the same condition in which it was when it was minted, showing the entirety of the original detail, as can be seen in Figure 2(c).

1.1.2 Miscellaneous

In order to appreciate the challenge of the task at hand, it is important to recognize a number of factors other than the condition which affect the appearance of a coin. These include die *centring*, surface metal changes (due to oxidation or other chemical reactions), and die wear.

Die centring refers to the degree to which the centre of the die coincides with the centre of the actual piece of metal against which it is struck to create the coin. A coin with poor centring may have salient design elements missing e.g. a part of the legend. An example of a somewhat poorly centred obverse can be seen in Figure 2(a).

Depending on the presence of different substances in a coin's environment (soil, air, etc.), the surface metal can change its colour and tone as it reacts with chemicals it is exposed to. Observe the difference in the tone of the coins in Figure 2.

Finally, it is worth noting that the appearance of a coin can be affected by die wear. Just as coins experience physical damage when handled and used, repeated use of a die in the minting process effects damage on the die. To a non-trained eye a coin minted with a worn die can seem identical to a worn coin minted with an intact die. However, a reasonably skilled (but not necessarily expert) numismatist can readily make a distinction, as subtler patterns of damage in the two cases are quite unlike one another. In addition, close inspection and the presence of oxidation or particles in ridges can be used for conclusive verification.

1.2. Previous work

Most early and some more recent attempts at the use of computer vision for coin analysis have concentrated on modern coins [13, 22, 23]. This is understandable considering that modern coins are machine produced and as such pose less of a challenge than ancient coins. Modern coins do not exhibit variation due to centring issues, shape, different depictions of semantically identical elements, etc. From the point of view of computer vision, two modern coins at the time of production are identical. This far more restricted problem setting allows for visual analysis to be conducted using holistic representations such as raw appearance [18] or edges [25], and off-the-shelf learning methods such as principal component analysis [18] or conventional neural networks [21]. However such approaches offer little promise in the context of ancient numismatics. Figure 1 illustrates the imposing challenge at hand.

Figure 1. Illustration of intra-class variation: each row shows coins (taken from our RIC-Hq data set and described in Section 3.1) with the same emperor on the obverse.

The existing work on computer vision based ancient coin analysis can be categorized by the specific problem addressed as well as by the technical methodology. As regards the former categorization, some prior work focuses on coin instance recognition i.e. the recognition of a specific coin rather than a coin type. This problem is of limited practical interest, its use being limited to such tasks as the identification of stolen coins or the detection of repeated entries in digital collections. Other works focus on coin type recognition, which is a far more difficult problem [26, 19, 3]. Most of these methods are local feature based, employing local feature descriptors such as SIFT [20] or SURF [11]. The reported performance of these methods has been rather disappointing and a major factor appears to be the loss of spatial, geometric relationship in the aforementioned representations. In an effort to overcome this limitation, a number of approaches which divide a coin into segments have been described [2]. These methods implicitly assume that coins have perfect centring, are registered accurately, and are nearly circular in shape. None of these assumptions are realistic [15, 12]. The sole method which does not make this set of assumptions builds meta-features which combine local appearance descriptors with their geometric relationships [3]. Though much more successful than the alterna-

tives, the performance of this method is still insufficiently good for most practical applications.

All of the aforementioned work shares the same limitation of little use of domain knowledge. In particular, the general layout of the key elements of Roman Imperial coins is generally fixed, save for few rare exceptions. Hence it makes sense to try to use this knowledge in analysis. The few attempts in the existing literature generally focus on the coin legend [6]. In broad terms this appears sensible as the legend carries a lot of information, much of which is shared with the coin's pictorial elements. For example, the obverse legend in almost all cases contains the name of the emperor depicted, and the reverse the name of the deity shown. The denarius of Antoninus Pius with Aequitas (goddess of justice and equality) in Figure 2 illustrates this well, the obverse legend being **ANTONINVS**AVG**PIVS**PPTRPCOSIII, and the reverse **AEQVITAS**AVG. However, in spite of this, methods such as that described by Arandjelović [6] offer little promise for practical use. The key reason for this lies in the fact that the legend, with its fine detail, is one of the first elements of the coin to experience damage and wear. Coins with clearly legible legends are generally expensive and rare, and thus of little interest to most collectors. They are also the easiest to identify, by the very nature of their good preservation, and hence do not represent the target data well. Consequently, this class of algorithms is not of interest in the present paper.

2. Proposed method

As we highlighted in the previous section, a stark methodological feature of nearly all methods on computer vision based analysis of ancient coins concerns their limited us of domain knowledge. Yet, there is an abundance of domain specific information that can be exploited. In particular, Roman Imperial coins exhibit regularity in terms of the presence and location of the key semantic elements. The first of these that a human expert focuses on is the portrait on the obverse of a coin, which (loosely speaking in numismatic terms) shows the issuing authority (emperor or empress) and hence narrows down the plausible rage of minting dates etc. In the present work we focus on this specific problem: that of recognizing the issuing authority based on its relief representation on a coin's obverse.

The task at hand can be seen as being a specific instance of face recognition which itself has received much research attention in the computer vision community. However, it is important to emphasise some key challenges that are shared and some which are specific to the two problems. Firstly, in both cases there exists intra-class appearance variation due to age, facial hair, and clothing, for example. Illumination changes also pose problems, albeit in different ways: while the surface of real human faces exhibits largely Lambertian reflectance properties [16], ancient coins are universally metallic. This means that unlike in the case of human faces the albedo of ancient coins is non-discriminative, as well as that colour [4] or edge based representations [9, 7] or features [5] are not reliable cues. Similarly, the assumption of discriminative content being primarily contained in high frequency signal bands, often made in face recognition research [8, 10], in the context of the task at hand holds poorly. Lastly, it is worth observing that Roman Imperial coins (unlike, say, Byzantine coins) always depict heads and busts in profile. The importance of this lies in the universally upheld finding of previous work that face recognition from profile is much more challenging than from frontal or semi-frontal poses [14].

2.1. Network architecture

On a coarse level, deep convolutional neural networks can be thought of as being comprised of layers of neurons, which by their connectedness structure can be categorized into different types. Most common of these include convolutional, pooling, and fully connected layers. In general, each layer has associated with it a different set of parameters, which together make up the entirety of connection weights and biases of the network. The parameters used specifically during the training procedure are referred to as hyperparameters.

The number of layer types with their individual parameters and the hyperparameters with their effect on training quality and speed make it difficult to choose a well performing architecture. For this reason, prior empirical evidence (that is, the performance of previously described architectures on structurally similar problems), and domain knowledge and insight into nature of the problem at hand, are crucial in guiding design choices. Our architecture is largely inspired by the finding of Simonyan and Zisserman [24] who demonstrated that a carefully crafted network built using few small (3×3), stacked kernels is superior to one comprising bigger kernels in terms of describability and computational cost. We make use of the rectified linear activation function (ReLU) which has been shown to improve convergence speed as well as generalization, compared to other nonlinear activation functions [17].

The network architecture we designed is made up of what we call convolutional blocks. Every convolutional block uses the same hyperparameters with the exception of the number of kernels, and is made up of two sets of convolutional layers, batch normalization, and rectified linear unit activation. Following a convolutional block a max-pooling operation in order to reduce the dimensionality of the transformed input.

The final architecture is made up of five consecutive convolutional blocks and max-pooling pairs. The number of filters is doubled after every pooling layer with the exception of the last layer. The output of the last pooling layer is

(a) Fine (b) Very fine (c) Extra fine

Figure 2. Specimens of a denarius of Antoninus Pius (RIC 61) from one of our data sets, RIC-Cond, described in Section 3.3.

flattened and then processed by 3 fully-connected layers of 4096 neurons followed by a soft-max output layer consisting of 83 outputs. Dropout is applied on the fully-connected layers except for the final output. The initial values of the weights and biases are drawn from a Gaussian distribution.

2.2. Training procedure and parameters

For the experiments in the present work the described network was implemented in TensorFlow [1] and trained on an Nvidia GTX 1080 graphics card. The time required for training was found to be approximately five hours. The hyperparameters were chosen by combining grid and random search strategies. Thereafter the main network parameters were optimized for using the cross-entropy error criterion, and stochastic gradient search with momentum using a mini batch size of 128, learning rate of 0.01, and momentum of 0.9. During training whenever the prediction accuracy on the training set failed to improve four times in a row, the learning rate was adjusted automatically by being halved. Training was deemed completed after four decreases in the learning rate. The final model, i.e. the corresponding network parameters, are selected as those which achieved the best prediction accuracy on the test set. This procedure is commonly referred to as 'early stopping' and is used to prevent overfitting.

3. Data

A major limitation of existing research on computer vision based ancient coin analysis concerns evaluation methodology. In particular, all of the published work in this area is characterized by experiments which by their design fail to control for pervasive confounding factors which affect the reported results. Many of the experiments were performed using extremely small data sets (thus failing to provide sufficient statistical significance), some use images acquired in highly controlled and uniform conditions (thus failing to provide insight on the generalizability of the evaluated methods), and none label, analyse, and control for the effects of different coin grades on the recognition performance. Guided by this failing of existing research, in our experiments we sought to address all of the aforementioned problems. Hence, we collected three new, large data sets, which we describe in detail next. For research purposes all of the data can be obtained freely from the authors upon request.

3.1. Data set 1 (RIC-Hq)

The data set that herein we refer to as RIC-Hq is the largest of the three data sets used in the present work. It is also the one with the most diversity of data: it spans the entire timespan of the Roman empire (from the rise of Octavian in 29 BC until the fall of the overthrow of Romulus in 476 AD), includes all denominations used in this period (sestertii, dupondii, asses, denarii, antoninianii, follēs, AE4s, AE3s, etc.), as well as coins of different conditions. Images in this data set were obtained by crawling web sites of reputable coin dealers and repositories of coins. The corpus comprises images of 29,807 coins which are highly heterogeneously distributed across different classes of interest (emperor). This is a consequence of the manner in which data were collected and the inherent differential in the scarcity of different emperors' coins – certain classes (usually those corresponding to emperors who ruled for prolonged periods of time, such as Antoninus Pius) contain many exemplars, while others very few (similarly, usually those which correspond to usurpers who ruled for brief periods of time, like Didius Julianus). Lastly, it is important to note that while the condition of coins in this data set is varied, the resolution of images and the quality of the conditions in which they were acquired are generally high, which is not surprising considering that they were produced for commercial purposes.

3.2. Data set 2 (RPC-Scan)

As we have mentioned earlier and as shall be elaborated shortly, one of the key methodological limitations of experiments in the existing literature on computer vision based ancient coin analysis concerns the uniformity of data in terms of its provenance, quality, and the manner of acquisition. In an effort to address this problem, we sought a second data set for our evaluation, which is very different than that described in the previous section. We achieved this goal by obtaining access to the collection of images of Roman Provincial coins in the collection of the Fitzwilliam Museum (Cambridge, UK) which houses one of the largest an-

(a)

(b)

Figure 3. Examples of two provincial coins of emperor Trajan (shown on the obverse in both cases) from one of our data sets, RPC-Scan, described in Section 3.2.

cient coin collections in the world. There are several key aspects in which this data differs from that in RIC-Hq. Firstly, the images were not acquired directly by digital cameras but are rather scans of photographs taken usually several decades ago. Hence the quality of data is affected both by the deterioration of the original photographs, and the artefacts formed during the scanning process. Moreover, being Roman Provincial rather than Imperial coins, the coins in this data set are very different stylistically (the corresponding dies were made by non-official mints, usually in distant places, and by individuals who had to rely on unreliable depictions of emperors as models), as well as in terms of their size, weight, and denomination from those in RIC-Hq. Comprising 19,164 coins, the size of RPC-Scan data set is somewhat smaller than that of RIC-Hq but vastly larger than any used in the existing literature (circa 3000 [3]).

3.3. Data set 3 (RIC-Cond)

One of the key methodological flaws in the evaluation of different algorithms described in the literature concerns the lack of consideration of the condition of the coins used for experiments. Given that the condition of a coin by definition affects the visibility and even the very presence of elements depicted on the coin, it is unsurprising that it is a major factor which governs the ease (or lack thereof) that a human experiences when attempting to identify a coin. Understanding the behaviour of different methods when presented with this challenge, and in particular the effects of both the condition of the query coin as well as of the distribution of coin conditions in the so-called gallery corpus, should be a crucial consideration in directing future research efforts.

At this point in time there does not exist a data set structured in a manner which allows for the analysis outlined above to be conducted: none of the corpora used in previous work can be readily adopted for use to this end, nor are there any other readily available sources, to the best of our knowledge. Hence we collected another novel data set for this purpose. In particular, we collected our data by searching for images of coins sold by well known ancient coin dealers. Having been put up for sale by reputable experts, the coins have been graded by professionals allowing us to associate reliable meta data with all images.

We collected 600 images in total. These represent 100 types of Roman Imperial denarii, with six exemplars for each type: two in fine condition, two in very fine, and two in extremely fine. The period covered by the coins included in the data set starts with the beginning of the Empire and the rule of Octavian in 27 BC and ends with the end of the rule of Philip II (Philip the Arab) in 249 AD when the denarius ceases to be used due to economic and political crises.

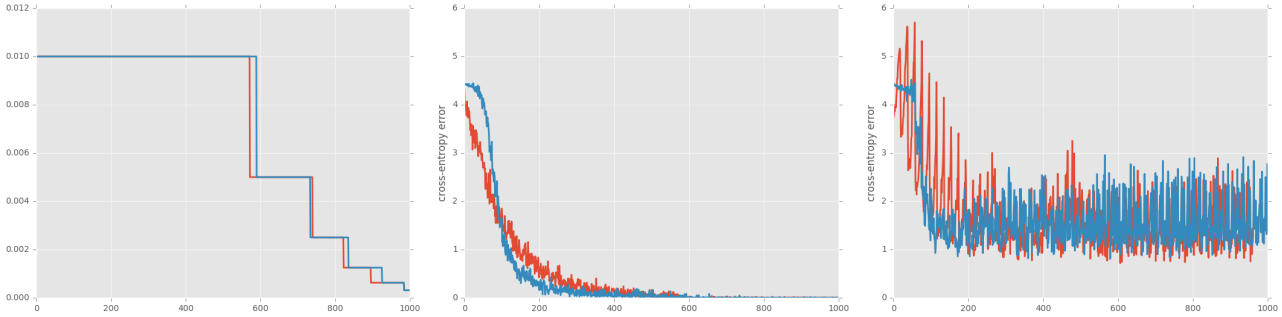
3.4. Training methodology

Our neural network was trained using a subset of the RIC-Hq data set. In particular, we split this data set into three equally sized subsets: training, test, and validation. The former two were used for all training in the present work, whereas the latter (together with RPC-Scan and RIC-Cond data) was used for the subsequent evaluation, as we will detail in the next section.

Considering the highly non-uniform distribution of exemplars across different classes (83 in total) in RIC-Hq we faced an interesting question of whether we should accept this as an inherent characteristic of data or if we should uniformly sample classes, "synthetically" effecting a balanced training data set. Our conclusion was that both strategies can be considered as reasonable under some circumstances; hence we conducted experiments using both approaches. In particular, accepting the original class imbalance is reasonable if one seeks to create a network with the best average classification performance. However, this view ignores an important practical consideration: coins of emperors with fewer exemplars are, by definition, rarer and thus usually much more expensive. Misclassifying an expensive coin can be therefore seen as effecting a greater penalty and synthetic rebalancing of classes can be seen as a way of implicitly producing the optimum network design from a utilitarian point of view.

3.5. Results and discussion

Considering the methodological flaws of experiments in the existing literature highlighted earlier, we took particular care to analyse the performance of each aspect of the pro-



(a) Learning rate (b) Training loss (c) Test loss

Figure 4. Temporal training characteristics. Red lines correspond to training using the original class size distribution (highly unbalanced) and blue to training using sampling that creates equally sized classes.

posed method in detail. We started by examining a range of relevant variables during the training stage.

3.5.1 Training

An illustration of typical training behaviour is presented in Figure 4. Firstly, relate the variation of the learning rate in Figure 4(a) to the previously described temporal adjustment of the same and observe that most of the training time is spent using the fastest learning rate and progressively less at each subsequent step when the learning rate is reduced. This behaviour is consistent with our theoretical prediction as the learning rate is adjusted, in effect, when the network reaches the vicinity of the global optimum to the extent of the precision possible with the current learning rate. From all four plots in the figure it can be seen that both training methodologies, namely using the original and synthetically balanced class sizes, exhibited similar behaviour on the global training scale. That being said, it is interesting to note the greater oscillation of the former in Figure 4(c) (test set loss). We expect that the likely explanation lies in random class size balance difference between the three sets: training, test, and validation.

3.5.2 Same data set classification

Our next step was to examine the performance of the proposed method on the same data set on which training was performed, RIC-Hq. We noted this before but to stress to point again, only the validation subset of RIC-Hq was used for this purpose i.e. training and test data were *only* used for training and were not included in the present experiment.

A summary of our results can be found in the second row Table 1. For completeness we also include the results for the training set, which are in the first row of the table. There are several important observations that should be made here. Firstly, note that already at rank-1 (the closest class is indeed the correct one) our algorithm achieves outstanding

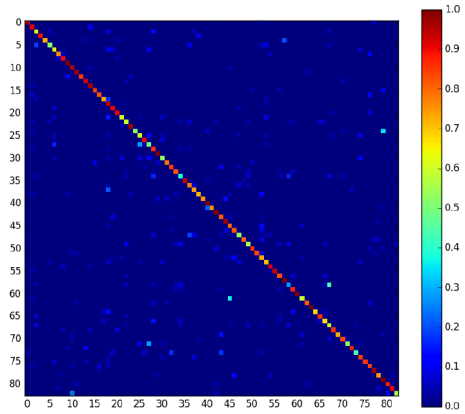


Figure 5. Confusion matrix for our first experiment (using the validation subset of RIC-Hq). Strong diagonal element dominance clearly shows that our algorithm makes correct classification decisions with a high decision confidence.

performance, exceeding 82% using the original class sizes, and approximately 80% when the classes are balanced (n.b. considering the variances, this difference is not statistically significant). This performance vastly exceeds that of all previous work [3] – applied on the same data the methods of [26] and [2] achieve rank-1 recognition of 3-4%, and that of [3] 12%. At rank-3 (the correct class is in the three closest as deemed by the algorithm) the results of our algorithm are little short of outstanding, exceeding 92%.

3.5.3 Generalizability

At several points in the manuscript we have highlighted the nearly universal use of high quality images by the existing methods, as well as the lack of variability in terms of how, when, and by whom experimental data was acquired. Usually all images were obtained from the same source. Hence our next step was to assess how our algorithm performs when applied on data from an entirely different provenance.

Table 1. Summary of experimental performance of our method on different data sets and using different training methodologies. Shown are the average recognition rates and the associated deviations across classes.

		Original exemplar per-class distribution		Uniform exemplar per-class distribution	
		Rank 1 (%)	Rank 3 (%)	Rank 1 (%)	Rank 3 (%)
RIC-Hq	Training data set	99.96 (±0.00)	100.00 (±0.00)	99.37 (±0.77)	99.94 (±0.09)
	Validation data set	82.53 (±2.66)	92.23 (±1.31)	79.79 (±2.80)	92.32 (±1.31)
RPC-Scan	Entire corpus	84.13 (±2.22)	94.11 (±1.01)	84.71 (±2.12)	93.92 (±1.00)
RIC-Cond	Fine	71.35 (±2.26)	84.40 (±2.98)	69.71 (±2.11)	84.12 (±2.26)
	Very fine	87.59 (±1.78)	97.24 (±0.00)	83.97 (±4.19)	96.55 (±1.26)
	Extremely fine	87.42 (±0.94)	97.35 (±0.94)	84.93 (±2.72)	95.70 (±1.27)

Figure 6. Class separation for a selection of particularly challenging coins (poor condition, non-uniform tone, etc.).

Using the same network as until now, trained on a subset of the RIC-Hq corpus, we examined its performance on the vastly different (see Section 3.2) RPC-Scan corpus.

A summary of our findings is shown in the third row of Table 1. An examination of the achieved recognition rates using different training methodologies and at different ranks of interest, provides clear evidence of remarkable generalizability of our method. The performance is virtually identical as on the validation set of the RIC-Hq corpus, both qualitatively (i.e. compared in relative terms with one another) and quantitatively. This conclusion is further corroborated when the results on the RIC-Cond corpus are considered (fourth to sixth rows of Table 1). We will discuss these results in more detail next.

3.5.4 Effects of grade

A major interest of ours and an important motivating factor behind the present work concerns the effect of coin grade on recognition performance, and the lack of systematic analysis of this effect in the existing literature. Our RIC-Cond data set was specifically collected and carefully annotated for this purpose.

The key results are summarized in rows four to six of Table 1. Specifically with regard to the effects of grade, there are several important conclusions that can be drawn. Firstly, as expected, classification performance was worst when our algorithm was applied on the lowest grade (F). A substantial improvement of approximately 16–18% can be observed for VF coins. Interestingly, no advantage of using extremely fine coins was found. That this is no coincidence is witnessed by the consistency between rank-1 and rank-3 matching, as well as across both training approaches. A possible explanation for this phenomenon may lie in the greater amount of exquisitely fine detail in coins so finely preserved. This detail often corresponds to idiosyncratic coin features specific to individual die engravers, rather than discriminative information in the context of coin classification.

Lastly, a quantitative comparison of the proposed method with a number of state of the art algorithms in the existing literature, is summarized in Table 2. The vastly superior performance of the former can be readily observed.

3.5.5 Additional analysis

Given the convincing evidence for outstanding overall performance of our approach in terms of its overall accuracy, robustness to data acquisition conditions, and coin grade, we sought to gain further and more detail insight into its behaviour. Firstly, we examined the confusion matrices associated with different data sets, in order to find out not only in what proportion of cases our algorithm correctly classified a coin, but also with what confidence. The confusion matrix from our first experiment (using the validation subset of RIC-Hq), shown in Figure 5, is typical, and its strong di-

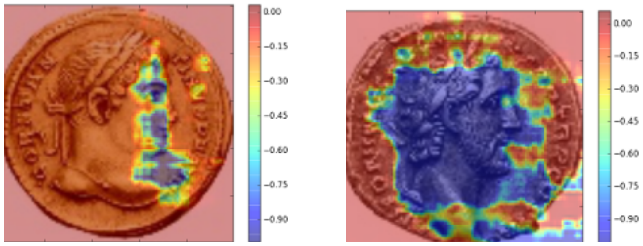
Figure 7. Examples of four different emperors, correctly recognized by our algorithm as different from on another but also as belonging to mutually the most similar classes. The emperors are, in order from left to right, Tetricus I, Tetricus II, Victorinus, and Quintillus.

Table 2. Comparison of the performance of the proposed method and the state of the art, on our RIC-Cond data set.

Method	Rank 1 recognition (%) F	VF	XF
Proposed	71.4	87.6	87.4
SIFT [26]	1.9	2.6	1.3
SIFT w/ wedge ($n = 4$) partitioning [2]	5.6	6.5	3.8
SIFT w/ sector ($n = 3$) partitioning [2]	6.1	8.8	6.7
SIFT w/ wedge & sector ($n = 4, m = 3$) partitioning [2]	3.1	3.2	2.1

agonal dominance clearly shows that our algorithm did not only make the correct classification decision in an outstandingly high proportion of cases but also that it did so with low risk. To confirm this further we additionally sought manually particularly challenging exemplars – namely, those in particularly poor condition or toning which hides detail – and examined our algorithm's performance on them in detail. Typical cases are illustrated in Figure 6, which shows a selection of coins, the top (correct) classification label with the corresponding confidence underneath each in green, and the next best matches (also with the corresponding confidences) in red. Outstanding separation between classes is readily apparent. A further case, magnified for greater clarity, is shown in Figure 7.

Lastly, using the occluder technique introduced by [27] and the occluder size 44×44 pixels, we examined the areas of coins images of different emperors which were learnt as having the greatest discriminative power. A representative illustration is shown in Figure 8, with bluer hues communicating higher importance. The results are extremely interesting and readily seen as meaningful by an expert. Figure 8(a) shows a coin of Constantine I (the Great) from the 4th century A.D. This period of economic crises and the overall decay of the Empire is characterized by a rather



(a) (b)

Figure 8. Salient image regions identified by our algorithm. Bluer hues indicate higher saliency.

generic depiction of emperors and little variation in the main elements of the portrait – the hair style, clothing on the bust, headwear (laurel wreath), etc. Hence the discriminative region is that which covers the main facial features. In contrast, the denarius of Antoninus Pius in Figure 8(b) was minted in the 2nd century during a period of increased prosperity, expansion, and economic development of the empire, and is associated with more detailed, artistic, and expressive coin engravings. The emperor's hair style itself is characteristic, the overall head shape, as well as the main facial features. Hence it is highly reassuring to observe that of all the aforementioned regions were correctly identified by our algorithm as salient for classification.

4. Summary and conclusions

In this work we addressed a major limitation of previous work on ancient coin based analysis which concerns the lack of use of domain specific information. In particular, exploiting the numismatic fact that Roman Imperial coins universally depict the ruling emperor on their obverses, we proposed a method which relies on profile face recognition for coin categorization. Our approach involves a carefully crafted deep convolutional network with a series of layers of different types, and comprised of small stacked kernels. Using by far the largest, most systematic, and detailed evaluation in the literature, our method is shown to outperform the state of the art by an order of magnitude.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv*, page 1603.04467, 2016. 4

[2] H. Anwar, S. Zambanini, and M. Kampel. Coarse-grained ancient coin classification using image-based reverse side motif recognition. *Machine Vision and Applications*, 26(2):295–304, 2015. 2, 6, 8

[3] O. Arandjelović. Automatic attribution of ancient Roman imperial coins. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1728–1734, 2010. 2, 5, 6

[4] O. Arandjelović. Colour invariants under a non-linear photometric camera model and their application to face recognition from video. *Pattern Recognition*, 45(7):2499–2509, 2012. 3

[5] O. Arandjelović. Object matching using boundary descriptors. *In Proc. British Machine Vision Conference*, 2012. DOI: 10.5244/C.26.85. 3

[6] O. Arandjelović. Reading ancient coins: automatically identifying denarii using obverse legend seeded retrieval. *In Proc. European Conference on Computer Vision*, 4:317–330, 2012. 3

[7] O. Arandjelović. Making the most of the self-quotient image in face recognition. *In Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, 2013. DOI: 10.1109/FG.2013.6553708. 3

[8] O. Arandjelović and R. Cipolla. Face set classification using maximally probable mutual modes. *In Proc. IAPR International Conference on Pattern Recognition*, pages 511–514, 2006. 3

[9] O. Arandjelović and R. Cipolla. A new look at filtering techniques for illumination invariance in automatic face recognition. *In Proc. IEEE International Conference on Automatic Face and Gesture Recognition*, pages 449–454, 2006. 3

[10] O. Arandjelović, R. I. Hammoud, and R. Cipolla. Thermal and reflectance based personal identification methodology in challenging variable illuminations. *Pattern Recognition*, 43(5):1801–1813, 2010. 3

[11] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008. 2

[12] B. Conn and O. Arandjelović. Towards computer vision based ancient coin recognition in the wild – automatic reliable image preprocessing and normalization. *In Proc. IEEE International Joint Conference on Neural Networks*, pages 1457–1464, 2017. 2

[13] P. Davidsson. Coin classification using a novel technique for learning characteristic decision trees by controlling the degree of generalization. *In Proc. International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pages 403–412, 1996. 2

[14] C. Ding, C. Xu, and D. Tao. Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing*, 24(3):980–993, 2015. 3

[15] C. Fare and O. Arandjelović. Ancient Roman coin retrieval: a new dataset and a systematic examination of the effects of coin grade. *In Proc. European Conference on Information Retrieval*, pages 410–423, 2017. 2

[16] A. S. Georghiades, D. J. Kriegman, and P. N. Belhumeur. Illumination cones for recognition under variable lighting: Faces. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 52–58, 1998. 3

[17] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. *In Proc. International Conference on Artificial Intelligence and Statistics*, 15(106):275, 2011. 3

[18] R. Huber, H. Ramoser, K. Mayer, H. Penz, and M. Rubik. Classification of coins using an eigenspace approach. *Pattern Recognition Letters*, 26(1):61–75, 2005. 2

[19] M. Kampel and M. Zaharieva. Recognizing ancient coins based on local features. *In Proc. International Symposium on Visual Computing*, 1:11–22, 2008. 2

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2003. 2

[21] Y. Mitsukura, M. Fukumi, and N. Akamatsu. Design and evaluation of neural networks for coin recognition by using GA and SA. *In Proc. IEEE International Joint Conference on Neural Networks*, 5:178–183, 2000. 2

[22] M. Nölle, H. Penz, M. Rubik, K. Mayer, I. Holländer, and R. Granec. Dagobert - a new coin recognition and sorting system. *In Proc. Internation Conference on Digital Image Computing*, 2003. 2

[23] X. Pan and L. Tougne. Topology-based character recognition method for coin date detection. *In Proc. IEEE International Conference on Image Analysis and Processing*, 2016. 2

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv*, page 1409.1556, 2014. 3

[25] L. van der Maaten and P. Boon. COIN-O-MATIC: A fast system for reliable coin classification. *In Proc. MUSCLE CIS Coin Recognition Competition Workshop*, pages 7–18, 2006. 2

[26] M. Zaharieva, M. Kampel, and S. Zambanini. Image based recognition of ancient coins. *In Proc. International Conference on Computer Analysis of Images and Patterns*, pages 547–554, 2007. 2, 6, 8

[27] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene CNNs. *arXiv*, page 1412.6856, 2014. 8