The Automatic Recognition of Ceramics from Only One Photo: The ArchAIDE App

Francesca Anichini[1][0000-0003-3813-9502], Nachum Dershowitz[2], Nevio Dubbini[1], Gabriele Gattiglia[1][0000-0002-4245-8939], Barak Itkin[2] and Lior Wolf [2]

[1] University of Pisa, MAPPA Lab, Via dei Mille 19, Pisa 56126, Italy
[2] School of Computer Science, Tel Aviv University, Ramat Aviv, Israel

Abstract. Pottery is of fundamental importance for understanding archaeological contexts. However, recognition of ceramics is still a manual, time-consuming activity, reliant on analogue catalogues created by specialists, held in archives and libraries. The ArchAIDE project worked to streamline, optimise, and economise the mundane aspects of these processes, using the latest automatic image recognition technology, while retaining key decision points necessary to create trusted results. The project has developed two complementary machine-learning tools to propose identifications based on images captured on site. One method relies on the shape of the fracture outline of a sherd; the other is based on decorative features. For the outline-identification tool, a novel deep-learning architecture was employed, integrating shape information from points along the inner and outer surfaces. The decoration classifier is based on relatively standard architectures used in image recognition. In both cases, training the classifiers required tackling challenges that arise when working with real-world archaeological data: the paucity of labelled data; extreme imbalance between instances of the different categories; and the need to avoid neglecting rare types and to take note of minute distinguishing features of some forms. The scarcity of training data was overcome by using synthetically-produced virtual potsherds and by employing multiple data-augmentation techniques. A novel way of training loss allowed us to overcome the problems caused by under-populated classes and non-homogeneous distribution of discriminative features.

Keywords: Ceramic identification, models, data, augmentation.

1 Introduction

Pottery is the most common type of excavated artefact, and its identification permits the understanding of the chronology, function, and importance of archaeological contexts. This identification is based on the archaeologist's domain knowledge and is usually made by matching potsherds to exemplars in catalogues of archaeological typologies. These catalogues contain, for each type, a standardised sketch of the complete vessel and sometimes a few photos. While not seeking to replace the knowledge and expertise of specialists, the ArchAIDE project worked to optimise and economise identification processes, developing a new system that streamlines the practice of pottery

recognition in archaeology, using the latest automated image recognition technology. At the same time, archaeologists remained at the heart of the decision-making process within the identification workflow. Specifically, ArchAIDE worked to support the essential classification and interpretation work of archaeologists (during both fieldwork and post-excavation analysis) with an innovative app for tablets and smartphones. The collaborative work of the archaeological and technical partners created a pipeline where potsherds are photographed, their characteristics compared against a trained neural network, and the results returned with suggested matches from a comparative collection with typical pottery types and characteristics. Once the correct type is identified, all relevant information for that type is linked to the new sherd and stored within a database that can be shared online.

The goals of the ArchAIDE project have been reported in (Wright and Gattiglia 2018; Anichini et al. 2020) and have been implemented through the creation of two distinct neural networks for shape-based and appearance-based recognition. The choice of the pottery classes, and, consequently, the catalogues to be used for the ArchAIDE project, was one of the main issues to be considered to create a system that requires a real-world implementation. The decision was made to choose four types: amphorae manufactured throughout the Roman world between the late 3rd century BCE and early 7th century CE; Roman Terra Sigillata manufactured in Italy, Spain and South Gaul between the 1st century BCE and the 3rd century CE; Majolica produced in Montelupo Fiorentino (Italy) between 14th and 18th centuries, and medieval and post-medieval Majolica from Barcelona and Valencia (Spain).

The set of tools the project developed addresses two scenarios: (i) when the pottery is undecorated, the identification relies on the geometry of the sherd; (ii) if visual patterns, such as colours and decorations, are preserved, classification is usually based on those, since they are more diagnostic than the shape of the sherd. Preliminary results on classification may be found in (Itkin et al., 2019).

Potential uses of computer vision and machine learning in archaeology were already proposed in (Van der Maaten et al., 2007) and applied to coin classification and to the retrieval of visually-similar glassware from a reference collection. Modern methods were used, for example, in (Orengo and Garcia-Molsosa, 2019) to detect and survey surface potsherds in high-resolution drone images. Detection and classification of whole pottery vessels in images by a prototype system called Arch-I-Scan is described in (Tuykin et al., 2018).

Much of the existing work on automated identification of sherds is based on 3D scanning or multi-view reconstruction technologies (Barreau et al. 2014; Calin et al. 2012; Kampel and Sablatnig, 2006; Karasik, 2010). However, the adoption of such methods is minimal due to the challenges of 3D acquisition in the field. The automatic analysis of profiles of potsherds has been studied using classical computer vision methods, but none is robust enough to be applied automatically on a varied set of excavated sherds. The problem of reconstruction from line drawing or sketches is classical (e.g. Malik, 1987; Tian et al., 2009; Yingze et al., 2009; Xu et al., 2014).

The complete processing chain for the two classifiers, shape-based and decoration-based, are sketched in Figures 7 and 8.

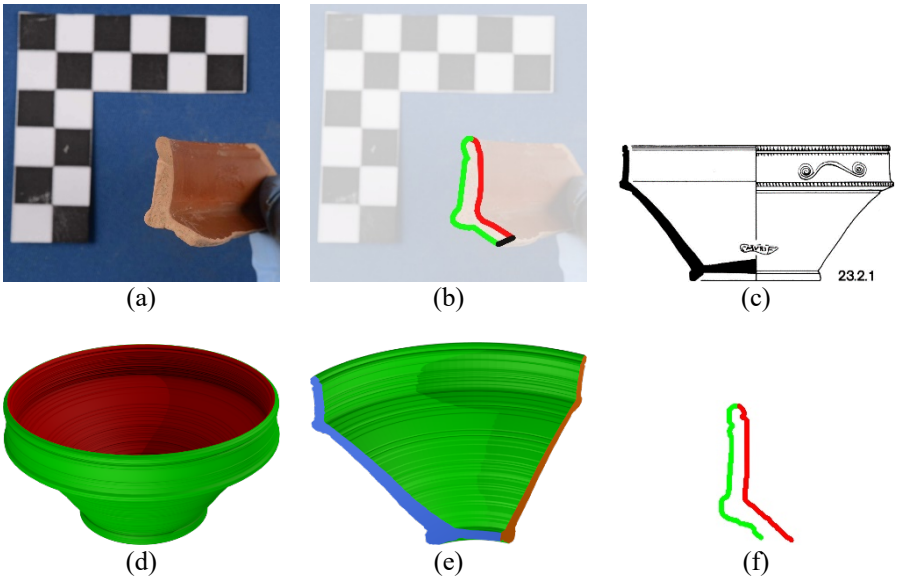


Fig. 1. An illustration of the archaeological data. (a) An image of a sherd positioned to show the fracture surface, with a reference scale ruler in the background. (b) A traced fracture outline overlaid over the source image. Green is the outer profile; red is the inner profile; black is for break lines that are ignored by the algorithm. (c) An archaeological sketch as it appears in a catalogue. One or more sketches define a class of pottery. (d) A 3D computer graphics vessel obtained by rotating the catalogue sketch. (e) A synthetic sherd obtained by breaking the 3D vessel. (f) A fracture outline obtained directly from the sketch, without the 3D reconstruction and shattering process.

1.1 Shape-Based Identification

Since our goal is aiding archaeologists in the field, we forgo multiple attempts to extract 3D geometry and rely on the 2D outline of the fracture surface of the sherd as the source of shape information. We tackle the task of classifying the outline of a potsherd based on a single image of it, as depicted in Figure 1(a). After marking the outline in a semi-automatic way and determining the scale using a ruler (Figure 1(b)), our AI-powered mobile app suggests an identification in the form of a list of archaeological types, ranked by their relevance to the pictured potsherd.

A major challenge in building the necessary AI tools is the lack of sufficient real-world samples to train neural networks. Furthermore, the variability in the dataset would still cover only a small fraction of the space of possible sherds. Instead, we define each class by one or more 2D sketches of the profile of the complete vessel; see Figure 1(c). Whereas the sketch describes the geometry of the profile of the entire vessel, an excavated sherd is a relatively small piece of the original, containing very limited information regarding the shape as a whole.

The outline of the fracture is a consequence of both the geometry of the pottery and the random breakage process. On the dataset side, we reconstruct the 3D pottery by rotating the profile of the vessel (Figure 1(d)) and shatter it to derive synthetic sherds (Figure 1(e)). We adopt a way to circumvent the computation overhead of 3D reconstruction and instead obtain the synthetic fracture surfaces (Figure 1(f)) efficiently.

To identify outlines, we train a network that supports the unique characteristics of archaeological outlines, including the need to separate between the inner outline and the outer outline of the sherd, the importance of the order of the points along the outline, the inherent noise in the tracing process, and the need to overcome sub-optimal data acquisition processes.

The architecture of our classifier relates to an emerging body of work, encoding inputs that are given as sets (Qi et al., 2017; Zaheer et al., 2017). It is similar to PointNet (Qi et al., 2017) in that it employs pooling in order to obtain a representation that is invariant to the order of the elements, following a local computation at each element. It has previously been shown (Qui et al., 2017; Zaheer et al., 2017) that, under mild conditions, such pooling is the only way to achieve this invariance. Recent works on shape classification include PointNet++ (Qi et al., 2017a), which employs local spatial relations, and PointCNN (Hua et al., 2018) which applies spatial information to group points prior to aligning them spatially to a grid where a convolution can be applied. While previous work has mostly been focused on the identification of 3D point clouds, we encode a 2D outline and benefit from the information that arises from the order of the points along the outline.

1.2 Decoration Identification

The case in which the pattern information on the face of the artefact is informative is much better addressed in the current computer vision literature. In this case, we employ a commonly used transfer-learning technique in which a neural network that was pre-trained to perform visual identification is adapted to the task at hand, using a relatively small archaeological training dataset.

1.3 Challenges and Results

Both in the shape and the decoration methods, we overcame a broad range of compounding challenges. These include: (1) the lack of real-world data to train on (shape) or a small one (decorations); (2) a partial view of the object that is obtained by a random breakage process, which presents large variability; (3) a large portion of the sherds, among both the synthetic training samples and the captured test samples, are almost entirely non-informative; (4) very similar classes, making the distinction more challenging and also causing ambiguity in the ground truth classification of the test data; and (5) a noisy acquisition process: an error-prone process for extracting the outline and obtaining scale from the real images (shape), variability in illumination (decorations).

In addition, to be used by experts, there is an acute need to optimise to fit considerations beyond accuracy. For example, most neural network loss measures would be

prone to sacrificing challenging classes to improve the average accuracy across all classes. However, a reference tool brings the most value when the identification is less obvious. To tackle the heterogeneous and unbalanced nature of the data, we train using a novel weighting technique that considers both the error of each ground truth class and false positives in each class. The reweighting scheme that we use addresses both the difficulty of correctly classifying a sample from a given class and the frequency of the current classification of a sample.

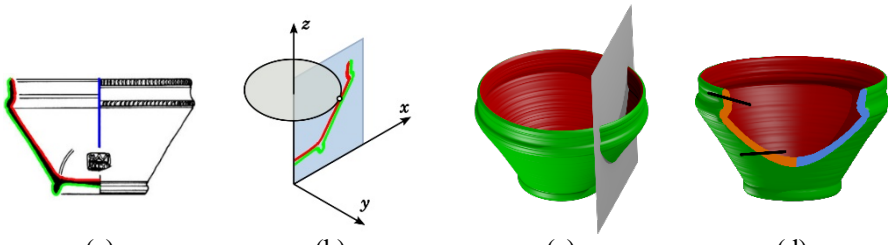
Our results demonstrate a relatively high recognition rate in the face of these challenges. The development was carried out in two phases to ensure the validity of our results. In the first, we developed the method on one dataset of potsherds of one specific family; in the second, the same method, with the same pipeline and (hyper-)parameters, was applied to three new datasets. With our Phase I dataset, out of 65 different classes, the tool can identify—based on images of sherds captured with a dedicated mobile app—almost 74% of the sherds within the top-10 results. With three additional datasets that were received after the completion of our research phase, without any tweaking of the pipeline, we reached 81%, 68%, and 60% top-10 accuracy for 65, 98, and 94 classes, respectively. Thus, our network may serve as the basis of a reliable reference tool for the use of archaeologists in the field, one that significantly narrows down the list of relevant classes to be considered for each sherd.

2 Shape-Based Identification

2.1 Synthetic Training Data

Generating high-quality data with as much similarity to real data as possible is crucial for our training. ArchAIDE process follows the steps described next to generate synthetic training data using the sketches extracted from the catalogues

Extraction of the profile from the sketch is done by tracing the edges of the profile of the vessel (the left half of Figure 1(c)). Handles, if present in the profile, are removed (Banterle et al., 2017). Finally, the scale is extracted from the ruler. A sample result can be seen in Figure 2(a). A 3D model can be obtained by rotating the profile around the vertical axis.



(a) (b) (c) (d)

Fig. 2. Sketch processing. (a) The processed sketch with the inner profile, outer profile, and rotation axis. (b) The rotation process. The inner and outer profiles are positioned for rotation around the rotation axis. (c) A cutting plane P through the 3D pottery. (d) The complete fracture face. In practice, only one of two sides (marked in orange and blue) is present in most excavated sherds. We further cut the top and bottom of the fracture, using two lines, to create a sherd with more realistic edges and size.

To generate a fracture directly from the profile, without reconstructing a computationally expensive 3D model, we imagine circles going around the vertical (z) axis, for each point in the profile (Figure 2(b)). We then generate a random 3D plane (Figure 2(c)), and compute the intersection of the plane with all the circles, connecting the intersection points from the circles along the profile to generate the fracture face (Figure 2(d)). To make the fracture shape more distinctive, we keep the random plane almost vertical (Figure 2). To add further realism to the generated fracture, after projecting the fracture back to 2D, we reduce its extent to match the dimensions of real potsherds; To do so, we cut the resulting polygon using two almost-horizontal lines (Figure 2(d)).

Since the drawings are scanned in high resolution to capture as many details as possible, artefacts resulting from the printing process may be visible (see Figure 3(a)) and reflected in the traced outline (Figure 3(b)). To avoid learning the artefacts, we simplify the outlines by sampling points randomly from each outline, limiting the number by resolution. When more points are needed (as training operates on a fixed number of points), we duplicate points as necessary. The network employs max-pooling, as detailed in Section 3.2, and seems to be able to overcome this inconsistency in sampling.

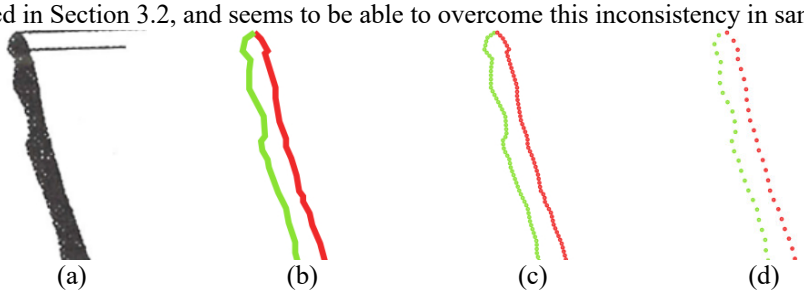


Fig. 2. Propagation of artefacts as a function of sampling resolution. (a) A scan of a drawing from the catalogue, depicting only the rim of a vessel and scanned at high resolution. Printing artefacts are clearly visible. (b) Accurate tracing of the drawing propagates some of the printing artefacts as rough edges. (c) Fixed-count sampling, matching the number of points required to achieve 2mm resolution on some of the larger potsherds. Due to the sample density, the tracing artefacts are still present. (d) A resolution-limited sampling, sampling every 2mm at the scale of the real pottery. Most artefacts are no longer visible.

When photographing potsherds, the fracture must be aligned with the image—where the sherd's vertical axis is aligned with the vertical axis of the image, and the fracture surface is kept parallel to the horizontal plane—to minimise distortions in the acquired fracture shape. Note that an archaeologist has no difficulty in approximating the vertical axis z since the ceramic manufacturing process creates shapes with dominant circles around z. The ability of the users to properly align the vertical axis (aligning both the vertical axis to the rotation axis and the fracture surface to the image plane) has been

verified in field trials. Despite the intuitive ability of users to align the fracture correctly, this alignment is inexact, since it is a manual process. For robustness, we simulate a small random 3D rotation on each fracture before projecting it onto a 2D outline.

Another concern with regard to data acquisition quality arises from the nature of the fieldwork. With one hand operating the camera and another hand holding the potsherd, a ruler that is used for inferring scale information is often left on the table and not held at the same distance as the fracture surface; see Figure 1(a). This seemingly small difference in distance from the camera, when combined with close-range photography, has been empirically shown to lead to scale computations that cause sherds to appear up to 50% larger than their actual size. To achieve robustness to this sort of issue, we also add a random scale factor.

2.2 Network Architecture

Our OutlineNet is based on PointNet with multiple improvements. Unlike PointCNN and Point-Net++, we do not attempt to cluster points together dynamically, but rather use the natural ordering of points along the outline for enriching the available information at each point with more than just its spatial location.

In our network, we supplement each point with two important pieces of information: (1) an annotation whether it is on the inside or outside, and (2) the angle of the outline at that point, which gives a rotation-invariant representation of the context around the point. The former is categorical; the latter is continuous. To combine them, we took an approach we called “group-hot” encoding; to represent d continuous values coupled with one categorical value with c options, we use a vector representing c groups of d values. To represent group i, we zero out the values of all but the i^{th} group and store the d values in that group.

Previous works construct hierarchies between points to encode spatial context for each point (Hau et al., 2018; Qi et al., 2017a). In our case, points are ordered, and we instead encode the immediate context around each point using angular information by considering, for every point, the cosine and sine of the angle formed at this point along the outline. Employing a point representation that incorporates both angle and location showed little to no benefit (compared to spatial information alone). Thus, we employ a multi-pathway architecture to enable learning separate features for spatial and angular information. We begin with separate branches of multilayer perceptrons (MLPs), one for angle data and one for location data. Their outputs are concatenated and fed into two perceptron layers. Max pooling is then performed over all points to obtain a global feature vector of the same size. Going through an additional MLP and a final softmax layer, we obtain output scores for the classes. All MLPs, except for the one producing the output score, employ ReLU activations.

3 Decoration-Based Identification

The drawings and the colours used to decorate pottery can be classified based on the usage of specific colours or their combination, by the type of patterns that are being

painted, by the areas that are being painted, and more. For appearance-based classification, our work was mainly carried out on the Majolica of Montelupo pottery. The data collection was led by the University of Pisa (UNIPI), using both existing images (from archaeological excavations, PhD theses, and more) and multiple photography campaigns. Most of the images were collected during the Autumn of 2017, with more than 8000 sherds being photographed, covering 67 genres with more than 20 sherds, many of which with more than 100 sherds. All the pictures have been classified by UNIPI archaeological staff.

Similar to other applications of computer vision in domains in which the data is relatively scarce, we rely on feature extraction from an existing neural network to the task at hand. As the base, we use a pre-trained version of the ResNet-50 network (He et al., 2016) trained on the ImageNet collection (Deng et al., 2009). The network operates on RGB colour images after these have all been resized to 224 x 224 pixels.

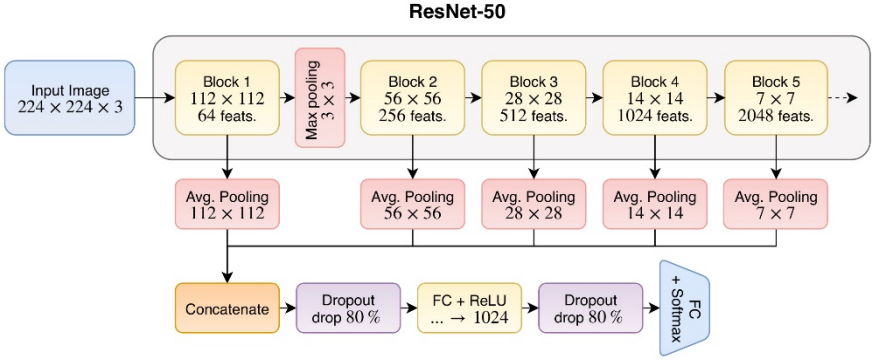


Fig. 3. The ResNet-based network for classifying potsherds by their appearance. The ResNet part of the network is frozen, and only the parts operating on the feature vectors are being trained. We use a significant dropout to reduce the overfitting that may occur with large feature vectors.

In order to utilise features at various levels of abstraction, we combine features from multiple levels of depth: while the lower levels encode colour and texture, the top layers encode complex patterns that are more related to the semantic content of the image. ResNet-50 is composed of a sequence of blocks, and we concatenate the features from blocks 2-5 in order to obtain one large feature vector, as can be seen in Figure 4. The feature maps from each block is a multidimensional map with a varying number of channels. Since we want to be position invariant, before this concatenation, we eliminate the spatial information by performing average pooling over the entire spatial extent of each channel, resulting in one vector of features from each block. To account for the different statistics of the features, we normalise each feature (in the concatenated vector) separately to have a mean of 0 and a variance of 1 on the training set.

The concatenated vector contains 3840 features. To this vector, we apply a dropout regularisation, a fully connected layer projecting to 1024 features, followed by a ReLU activation, a second dropout, and a projection to the number of classes followed by a softmax operator. Both dropout layers employ a high level of drop (80%) in order to increase robustness and decrease reliance on specific features. During training, we fix

the parameters of the ResNet layers that extract the features and only train the parameters of the fully connected layers on top of the features.

To fit the images to the expected input dimensions of the ResNet model, we scale them to 224 pixels (along the shorter axis) and crop them (equally on each side of the longer axis) to obtain a 224x224 image. To train our network to work with varying amounts of decorations/background inside the image, we enrich the original image dataset by adding augmented versions of each image: for each image, we scale it to four different sizes; on each scaled image, we create three flipped versions (unflipped, horizontally flipped and vertically flipped); we crop all of those images, leaving just the centre square. Thus, from each image, we create 12 images that can go into the neural network, increasing the dataset size from around 8000 images to about 100,000 images.

In our initial experiments, varying illumination was the most challenging factor in identification. To solve this lack of robustness, we simulate different white balance results and various brightness and contrast adjustments. This was applied during the generation of the training dataset by multiplying the luminosity ("brightness") of all the pixels within each image, using a randomised factor to simulate different lighting conditions. To compensate for different white balance setups, we additionally apply a similar random multiplicative factor to each channel in the image; that is, we multiply each of the red/green/blue channels, by a separate random constant factor, to change the ratio between colours in the image.

In addition, the imaging conditioned (background, ruler) varied considerably between the collection campaigns and the other sources, leading to an inherent bias (as each campaign had different kinds of pottery), as can be seen in Figure 5. To overcome this, we extract the foreground of the training images automatically. During testing, the GrabCut algorithm (Rother et al., 2004) was used to extract the relevant image part.

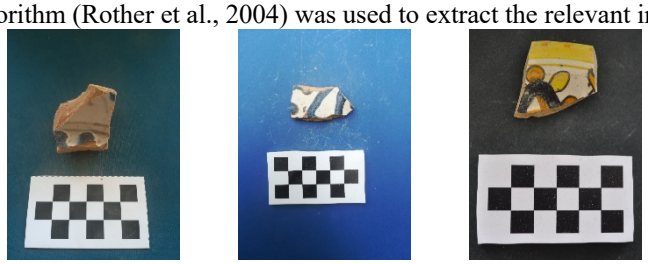


Fig. 4. Three typical images captured during the photography campaign.

4 Loss Reweighting

Most common techniques for combating low-classification accuracy introduce weights on the loss expressions of individual samples, with higher weights assigned to inputs from classes with low accuracy. While the rationale is clear, there is no guarantee that it will make the classifier learn anything "meaningful" about the classes. To mitigate this issue, we employ a new loss function, dubbed CareLoss, which weights samples not just by their true label but also by their predicted label. For each sample, the loss has one weight by the true label (assigning higher weights for classes with low

accuracy) and another weight by the predicted label, assigning higher weights to misclassifications. The second weight is aimed at tackling an increase in accuracy, which is accompanied by an increase in the number of false positives.

As it turns out, the new loss function not only increases the uniformity of the accuracy among the classes but also increases the overall performance on the test set. We attribute this to the fact that during testing, the same types of confusions that occur in the training data are likely to occur, only more frequently. This loss function was successfully applied to both appearance-based and shape-based identification.

5 Experiments

5.1 Shape-based identification

The development of the reference tool was planned as a two-phase process, where we first develop the classification algorithm on one dataset and then validate it on multiple other datasets for different types of pottery. Separation of datasets enables avoiding overfitting due to multiple hypotheses testing, thus enabling better confidence in our results. The dataset used in the first phase is made of 435 sketches of Terra Sigillata Italica (TSI), grouped into 65 standardised top-level classes, as defined in the Conspectus catalogue (Ettlinger et al., 2002). From these drawings, we generated class-balanced synthetic data, while reserving the outlines of the real-world sherds, to be used exclusively for testing. The real-world outlines were extracted from images collected across Europe using a dedicated mobile app.

To obtain the outlines, the user taps with their finger on a touch screen, marking the points of the outline and annotating these with side information (inner or outer outline). The manual annotations result in coarse polygons, thus making the dataset more challenging due to lack of fine details, and inaccuracies resulting from a touch-based input. The real-world test dataset contains 240 extracted outlines from 29 different top-level classes. Nevertheless, we train our classifier on all 65 classes. When training our model, OutlineNet's real-world top-2 classification rate was 1.5 times the top-1 classification rate. This indicates that the classes are easily confused. Ablation experiments showed that separation of inner and outer outlines, angle information, group-hot encoding, and adaptive sampling each add to the overall top-K performance, even when changes in the top-1 accuracy were small. Similarly, augmentation also contributed to the top-K result, without significant impact on the top-1 accuracy. A plausible reason is that all these modifications to the model and training, are less meaningful for samples that are carefully collected and informative, and mainly impact the accuracy of the lower-quality samples.

This befits its use as a reference tool for domain experts who would be happy to consider a short list of results as part of the mandatory expert verification but would be discouraged to use a tool that often completely omits the correct result.

Following the first phase development on the Terra Sigillata Italica (TSI) dataset, we obtained three additional datasets. The first was an additional TSI dataset, collected with the aid of the app. It includes the outlines of a further 96 actual sherds not included

in our previous dataset of real data and belonging to 11 classes previously unseen. Two additional datasets, Terra Sigillata Hispanica (TSH) and Terra Sigillata South Gaulish (TSSG) were added. These also belong to Terra Sigillata pottery but has different geographical origins, manufacturers, set of classes, and typology. (There is no intersection in classes between TSI, TSH, and TSSG.)

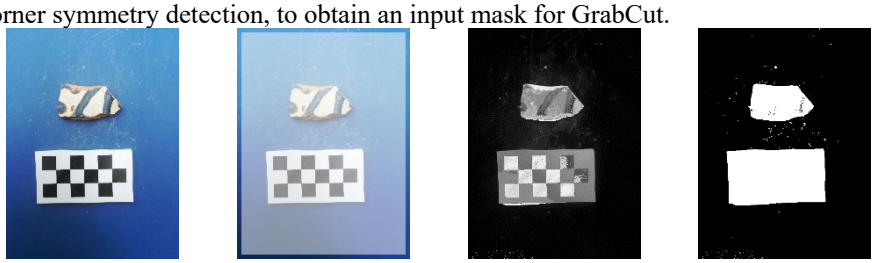
On the new TSI test set, using the same model from phase I (without any retraining/adaptations), the accuracy values obtained are even better than the phase I dataset. Additionally, for the datasets using new typologies, similar or better accuracies (measured relative to the number of classes) were obtained using exactly the same training method, without any adaptations.

5.2 Decoration-based identification

Experiments with decoration-based identification were carried out mainly with Majolica of Montelupo pottery, also in a two-stage process. In the first stage, the model was trained on a dataset, and while demonstrating promising results, evaluations made using real potsherds captured in varying conditions (and not using pictures from the dataset), demonstrated poor robustness of the classification process.

After a thorough ablation process, and as mentioned in section 4, the key differentiators in the classification results were found to be varying backgrounds and varying lighting conditions. While varying lighting conditions could be simulated during training (by augmenting the image), removing the background and ruler from the image (as these are correlated to specific classes thus generating a bias) was more challenging. After integrating an interactive extraction algorithm (GrabCut) to be used in the app, going back to extract the background from thousands of images in the dataset was not a reasonable effort.

After experimenting with multiple options, we developed a heuristic to fill the interactive role that is traditionally required in GrabCut. First, we collect values along the edges of the image (Figure 6(b)), and then compute the colour-distance of all pixels in the image from the nearest edge colour (Figure 6(c)). Applying dynamic thresholding to obtain two islands (Figure 6(d)), we can now remove the ruler island via simple corner symmetry detection, to obtain an input mask for GrabCut.



(a) (b) (c) (d)

Fig. 5. The process to generate the masks for GrabCut. (a) The input image, (b) the edge pixels we sample for their colours, (c) a measure of the colour distance of each pixel from the nearest edge pixel, (d) a threshold to obtain two white patches (sherd and ruler).

This heuristic for background extraction worked well for most, but not all images. Nevertheless, retraining the model with the background removed automatically and lighting augmentation, produced more robust results significantly in the face of varying photography conditions.

6 Results

6.1 Shape-based identification

The evaluation of the shape-based identification was done both on the captured real-world data (used in the testing phase), and also in an end-to-end fashion, with users capturing new photos, annotating them, and using the classification algorithm.

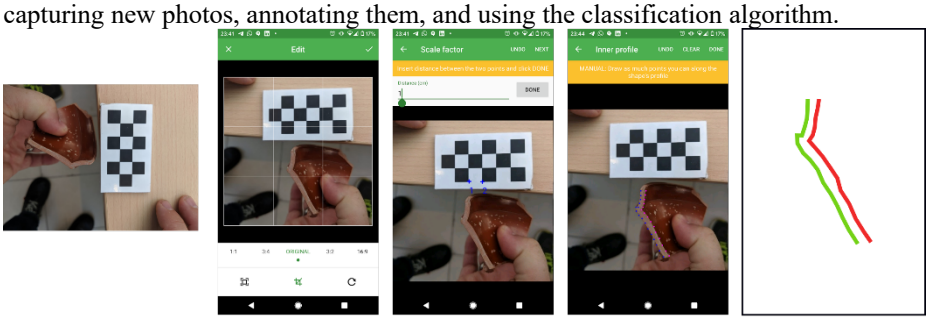


Fig. 6. The process of obtaining the outlines for the classification is described in the above figure. We start with an input picture which is captured using a smartphone. The image is then cropped and aligned to align the rotation axis with the vertical axis of the image. Afterwards, the scale is extracted by marking the physical distance between two points on the ruler. Finally, the inner and outer outlines are annotated by tapping on the screen to mark the outline points. The resulting shape is then classified by our model.

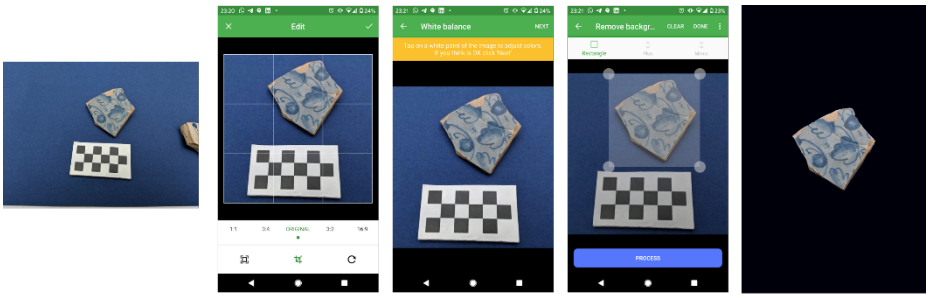
The end-to-end evaluation was done using 381 different pictures of sherds of TSI, taken from 42 (out of 65) different types. Most images were captured with a smartphone or a tablet (as would be the case on the field), with only 25 pictures using a regular camera. The average mobile-app top-5 accuracy is 50.8%, and the top-1 accuracy is 18.9%. This is slightly lower than 22.0% top-1 accuracy and 57.9% top-5 accuracy reported in our evaluation, but these results are still good and usable for archaeologists.

The results reported on the testing data, evaluated in a broader set of images, across multiple datasets, are reported below:

Accuracy	TSI (#1)	TSI (#2)	TSH	TSSG
Top-1	22.0%	30.5%	27.6%	14.5%

Top-2	32.7%	43.6%	40.6%	25.0%
Top-5	57.9%	62.8%	58.4%	41.9%

6.2 Decoration-based identification



Input Picture | *Crop & Align* | *White Balance* | *Interactive Background Removal* | *Sherd without background (image data)*

Fig. 7. The process of obtaining the potsherd images for the classification is described in the above figure. We start with an input picture which is captured using a smartphone. The image is then cropped and aligned to remove other potsherds that might be visible in the same image. Afterwards, white balance is performed to correct the image colours, and finally, interactive background extraction is performed using the (interactive) GrabCut algorithm.

The evaluation of the decoration identification method was done on both the mobile and desktop versions, including testing of different lighting conditions (as these were a key factor in the classification results for the first version). The results for the classification are reported below:

Accuracy	Mobile Performance	Desktop Performance
Top-1	55.2%	51.0%
Top-5	83.8%	77.2%

The analysis was conducted on 49 different genres (out of 84) with more than 700 images taken on mobile devices (phones and tablets) and more than 120 taken with a camera and classified in the desktop app.

Further results show that the accuracy of appearance-based recognition, on both mobile devices and desktop, is not related to the light type, being approximately equal with artificial and natural light.

7 Discuss

Several lessons were learned.

A first key lesson was that ArchAIDE has demonstrated the potential of using automated image recognition to identify archaeological pottery, even if at first glance, the results may appear unbalanced in comparison with the large-scale data capturing and the functionality of the algorithms. ArchAIDE has implemented original techniques for data generation and augmentation, for the weighting of samples and for line-based shape recognition. These are directed towards specific challenges in the classification of highly complex data such as pottery fragments, which include 3D and 2D information and multiple factors that complicate a homogeneous recording. Although the classification validation results might not look particularly striking in comparison with other artificial intelligence application in archaeology, these are impressive given the difficulty of the task at hand and represent a significant advance on the road to automated pottery classification. The achieved level of accuracy has been calculated on the full number of types or decorations known for a pottery class, which also contain very uncommon types or decorations. On the contrary, the level of satisfaction of the archaeologists who used the application on the field is higher than the raw number of the accuracy results. This is because the system can recognise all the more common types archaeologists found. In the case of Majolica of Montelupo, for example, the algorithm identifies with difficulty only the decorations realised in a few specimens for a richer client. These are not found in archaeological excavations, they are conserved in a museum, but their decoration is listed within catalogues. In this case, 83% of the accuracy of appearance-based recognition represent the full totality of archaeological finds. Moreover, ArchAIDE has also shown that it may be used for a variety of pottery types if the necessary comparative data can be gathered (and potentially other artefact types as well), as virtually all pottery identification relies on recognition based on either the shape or decorative elements of a vessel (or both).

The second lesson was the amount of training data necessary for the image recognition algorithm to return useful results. In our case, multiple photo campaigns were conducted across the life of the project to produce a complete dataset of images for all the ceramic classes under study. The photo campaigns aimed to provide a sufficient number of images to train the algorithms for both the appearance-based (Majolica of Montelupo and Majolica from Barcelona) and the shape-based image recognition neural network (Roman amphorae, Terra Sigillata Italica, Hispanica and South Gaulish). As not all types were stored in a single site, it was necessary to access multiple resources involving more than 30 different institutions in Italy, Spain, and Austria. To train the shape-based neural network was essential to take diagnostic photos of sherd profiles, so detailed guidelines were prepared for use by the consortium partners and project associates. Finding, classifying, photographing and creating digital storage for the necessary sherds was very time-consuming, as images of at least ten different sherds for every type were needed to provide enough training information for the algorithm. It became apparent that not every top-level type and sub-type could be represented. In some instances, this was because the type was rare, or because sherds of different types were mixed when stored, and it was challenging to locate them. This task is a challenge

across all forms of pottery studies, not just for a digital application like ArchAIDE. Overall, 3498 sherds were photographed for training the shape-based recognition model. For appearance-based recognition, using every image where the decoration was visible, it was possible to collect photos taken for different purposes, e.g. graduate or PhD theses, archaeological excavations, etc. In these cases, photos were collected, classified, tagged, and stored based on the genres of decoration to which they belonged. A larger corpus of pictures was collected through photo campaigns in Italy and Spain. A total of 13,676 photos were obtained. This resulted in far more time and effort spent on digitising the paper catalogues and undertaking the enormous photo campaigns to capture the necessary primary data. This effort helped partners understand the importance of working together if the humanities wish to take advantage of the many machine-learning methods now available. Datasets are small, fragmented, and rarely optimised for machine-learning applications.

The third lesson was that it was not reasonable to design an image recognition system that could identify pottery using *both* decoration-based and shape-based characteristics. It took considerable effort and discussion, but it became clear that it was necessary to separate them, developing two different algorithms. From an archaeological point of view, this does not represent a problem. If needed, ceramic classes for which both shape data and appearance data are available can be recognised using the two different classifiers in order to obtain more detailed results. Moreover, the project represents a proof of concept, and new experiments could be conducted with other ceramic classes. This choice allowed a creative outcome, as separating shape-based recognition allowed the 3D models to be used to create desperately needed training data. By "breaking" the models into "virtual sherds" and using the sherds to train the shape-based image recognition algorithm, the accuracy rate was increased to an acceptable level.

Finally, archaeological classification is not made purely based on the shape or decoration. Additional domain expertise, which is not currently captured in our scheme, enables the archaeologist to filter out some classes based on the location of the findings, other findings in the excavation site, and various other considerations. This by itself is not a technological limitation, as this sort of filtering can be implemented on top of the class ranking predicted by our reference tool. However, it means that the gap in the ability to distinguish potsherds based on their shape or decoration, vs human archaeologists, is probably much lower than the error rates of our method.

Another reason to believe that the error rates are probably inflated is that the labelling of individual potsherds is gathered from accepted labelling that is documented in catalogues and established collections. However, in some cases, the exact provenance of the assignment has been lost, and the ground truth classification is likely to contain mistakes.

To tackle a real-world cross-modality matching problem that presents a large set of compounding challenges, we conceived of multiple innovations, including the design of novel data generation techniques, a new shape representation scheme, and an original reweighting method. Our work also provides—beyond various technical novelties and a working application—a case study of deep learning applied to real-world data in a

situation where most of the conventional assumptions are grossly violated, and the reality gap ("sim2real domain shift") is wide, and the simulation must be done with significant care.

The method described in this paper is already deployed in the field as the main part of an archaeological reference tool. The source code, models and data are already made public.

Acknowledgements. This research was supported by the EU Horizon 2020 grant agreement No. 693548. We thank all the members of the ArchAIDE (archaide.eu) team.

References

Anichini, F., Banterle, F., Buxeda, I Garrigós J., Callieri, M., Dershowitz, N, Diaz Lucendo, D., Evans, T., Gattiglia, G., Gualandi, M.L, Hervas, M.A., Itkin, B., Madrid I Fernandez, M., Miguel Gascón, E., Remmy, M., Richards, J., Scopigno, R., Vila, L., Wolf, L., Wright, H., Zallocco, M.: Developing the ArchAIDE application: A digital workflow for identifying, organising and sharing archaeological pottery using automated image recognition, Internet Archaeology 52. https://doi.org/10.11141/ia.52.7 (2020)

Banterle, F., Itkin, B, Dellepiane, M., Wolf, L., Callieri, M., Dershowitz, N. and Scopigno, R.: VASESKETCH: Automatic 3D representation of pottery from paper catalog drawings. In: 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 9-15 Nov 2017, Kyoto, Japan. 683–90. https://doi.org/10.1109/ICDAR.2017.394 (2017).

Barreau, J-B, Nicolas, T., Bruniaux, G., Petit, E., Petit, Q., Gaugne, R., Gouranton, V.: Ceramics fragments digitisation by photogrammetry, reconstructions and applications. In: International Conference on Cultural Heritage, EuroMed, Lemessos, Cyprus (2014).

Calin, N., Popescu, S., Popescu, D., Mateescu, R.: Using reverse engineering in archaeology: Ceramic pottery reconstruction. Journal of Automation, Mobile Robotics and Intelligent Systems 6(2), 55–59 (2012).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, pp. 248-255, doi: 10.1109/CVPR.2009.5206848 (2009).

Ettlinger, E., Römisch-Germanische Kommission Des Deutschen Archäologischen Instituts zu Frankfurt: Conspectus formarum terrae sigillatae Italico modo confectae. Habelt, Bonn (2002).

He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778 (2016).

Hua, B.S., Tran, M. K., Yeung, S.K.: Pointwise convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 984-993 (2018).

Itkin, B., Wolf, L., Dershowitz, N.: Computational ceramicology. arXiv: 1911.09960 (2019).

Kampel M., Sablatnig R.: 3D data retrieval of archaeological pottery. In: Zha H., Pan Z., Thwaites H., Addison A.C., Forte M. (eds) Interactive Technologies and Sociotechnical Systems. VSMM 2006. Lecture Notes in Computer Science, vol 4270. Springer, Berlin, pp. 387–395 (2006).

Karasik, A.: A complete, automatic procedure for pottery documentation and analysis. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, pp. 29-34, doi: 10.1109/CVPRW.2010.5543563 (2010).

van der Maaten, L.J.P., Boon, P.J., Paijmans, J.J., Lange, A.G., Postma, E.O.: Computer vision and machine learning for archaeology. In J.T. Clark and M. Hagemeister (eds.) Digital Discovery. Exploring New Frontiers in Human Heritage. Computer Applications and Quantitative Methods in Archaeology. Archaeolingua, Budapest. https://lvdmaaten.github.io/publications/papers/CAA_2006.pdf (2007).

Malik, J.: Interpreting line drawings of curved objects. International Journal of Computer Vision 1(1), 73–103 (1987).

Orengo, H.A., Garcia-Molsosa, A.: A brave new world for archaeological survey: automated machine learning-based potsherd detection using high-resolution drone imagery. Journal of Archaeological Science, 112: 105013 (2019).

Qi, C.R., Su, H., Kaichun, M., Guibas, L.J.: PointNet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, pp. 77-85, doi: 10.1109/CVPR.2017.16 (2017).

Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems, pp. 5099-5108 (2017a).

Rother, C., Kolmogorov, V., Blake, A.: "GrabCut": Interactive foreground extraction using iterated graph cuts. In: ACM SIGGRAPH 2004 Papers (SIGGRAPH '04). Association for Computing Machinery, New York, NY, USA, pp. 309–314. DOI:https://doi.org/10.1145/1186562.1015720 (2004).

Tian, C., Masry, M.A., Lipson, H.: Physical sketching: Reconstruction and analysis of 3D objects from freehand sketches. Computer-Aided Design, 41(3), 147–158, doi:10.1016/j.cad.2009.02.002 (2009).

Tyukin, I., Sofeikov, K., Levesley, J., Gorban, A.N., Allison, P., Cooper, N.J. Exploring automated pottery identification [Arch-I-Scan], Internet Archaeology, 50. https://doi.org/10.11141/ia.50.11 (2018).

Wright, H., Gattiglia, G.: ArchAIDE: Archaeological automatic interpretation and documentation of ceramics. In: Proceedings of the Workshop on Cultural Informatics Research and Applications, colocated with the International Conference on Digital Heritage, Nicosia, Cyprus, November 2018: 60-65 (2018).

Xu, B., Chang, W., Sheffer, A., Bousseau, A., McCrae, J., Singh, K.: True2Form: 3D curve networks from 2D sketches via selective regularisation. ACM Transactions on Graphics 33(4), Article 131. DOI:https://doi.org/10.1145/2601097.2601128 (2014).

Yingze, W., Chen, Y., Liu, J., Tang, X.: 3D reconstruction of curved objects from single 2D line drawings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, pp. 1834-1841, doi: 10.1109/CVPR.2009.5206841 (2009).

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R., Smola, A.J.: Deep sets. In: Neural Information Processing Systems, pp. 3394–3404 (2017).