https://doi.org/10.2352/EI.2022.34.3.MOBMU-362

Recognition of Objects from Looted Excavations by Smartphone App and Deep Learning

Waldemar Berchtold, Huajian Liu, Simon Bugert, York Yannikos, Jingcun Wang, Julian Heeger, Martin Steinebach, Marco Frühwein; Fraunhofer Institute for Secure Information Technology SIT / ATHENE; Darmstadt, Hesse/Germany

Abstract

In this paper, we present a development for recognizing objects from looted excavations. Experts with required expertise are not always available where an archaeological object needs to be assessed for import, export or trade. For this purpose, we developed a smartphone app that can provide on-site assistance in the initial assessment of archaeological objects. The app sends captured images to a server for recognition and receives results with similar objects and their metadata along with an associated probability. A user can thus use these information to infer the provenance of the photographed object. To this end, a deep learning based solution was developed to identify archaeological objects, including a classifier trained using transfer learning and an image matching scheme based on deep convolutional neural networks (CNN) features. The developed application will be tested by law enforcement agencies with a total of 15 smartphones for six months starting in early October.

Introduction

The trade with antique objects from looted excavations is a lucrative source of income for some criminal organizations [8] [9] [10]. The transport of these objects often bypasses customs and law enforcement agencies.

The subject of this work is to provide a tool to recognize antiquities that have been recovered through looted excavations and are now to be sold. This can happen either because an object is deemed suspicious due to its similarity to known objects or because written information about the object does not match the information obtained to similar known objects.

Currently, trained experts are needed to be able to classify an object into the region and time period of its origin. The experts are not always available when needed for classification in a specific case. To counter this, we have developed an application consisting of a mobile app and a server. This application is intended to support authorities in obtaining an initial assessment of the geological region and which period the object most likely originated from. Depending on these two parameters, an object is either tradable, or the object is retained for further investigation.

The user of the app can take up to six images from different directions and send them to the server. The server uses a deep trained network to extract features that identify the style and thus an era and region, summarized as *metadata*. The server sends the most similar images from the database with the metadata to the app. Based on the similar images and the related metadata, the app user can decide what to do further with the object.

The underlying assumption of the approach is that the style, which was characteristic for this epoch and a region, can be extracted on the basis of the characteristics. If the style matches, the metadata for a newly photographed object can be assigned based on the style. For this approach no pre-trained network like e.g. ResNet can be used for the classification, because here no object recognition in the classical sense helps us. The reason is that we are not interested in whether a vase, a bowl, a plate or a cup was taken in the image, but in which collection an object belongs to. A collection consists of various types of objects, such as vases, coins, plates, potsherds, paintings, etc., all of which can be assigned to a specific region and epoch.

A two-step classification, which first classifies the object type with a pre-trained model, e.g. with ResNet, and then classifies the two metadata region and epoch in a second step, is not chosen for two reasons. First, our classes in the training data would be too small and thus worse results are to be expected. Second, the style is usually the same across different object types for a region and epoch which the model most likely cannot learn sufficiently well in this two-step process.

For this reason, a classifier is trained in the scope of this work to classify by collections. In order for an app user to be able to classify the result, we decided to perform an image matching within the classified collections after the classification and display the most similar objects from this collection including the metadata. Alternatively, an algorithm for interpreting the classifier, such as LIME, could have been used here.

Similarity vs identity

The method presented here relies on a technology that makes it possible to identify objects similar to the object under investigation and thus provide metadata of these objects. Here, a brief consideration of the basic technical choices and alternatives will be made in order to be able to place it in the context of different methods of image recognition.

Classifying three-dimensional objects is a special challenge, because during the examinations it is not known from which angle the reference objects were photographed. Therefore, the app provides for photographing the objects from several angles and asks the user from which perspective the image was taken. Two-dimensional objects, usually fonts or pictures, have this problem only to a limited extent. Although the object may be photographed from different angles, the resulting distortion can be compensated relatively easily.

Since it can be assumed from the usage scenario that objects to be investigated are not already recorded in the data set. Since we are dealing with looted excavations, it must be assumed that the exact object being investigated was not previously known and that there are no catalog entries or the like for it.

Therefore, a recognition in KIKu cannot focus on the recognition of identical objects, but must be limited to a recognition

of similar objects. This has direct effects on the applicable procedures. The following approaches are generally known: cryptographic hash methods, robust hash methods, feature recognition and image classification.

Cryptographic hash methods Only identical digital objects can be recognized with them. They are therefore not suitable for a photography-based solution, since even a photograph of an image with a cryptographic hash would not be recognized, because any kind of unavoidable image noise would change the hash.

Robust hash methods These methods are suitable for recognizing objects. They are especially common in the image domain, but are also known for videos, audio data or texts. With them, in contrast to a cryptographic hash, an image could be recognized even if it has been altered to some extent by noise. With them, images can be recognized despite changes. Nevertheless, robust hashing methods are not suitable for recognizing similar objects, since their design is to avoid confusing similar objects.

Feature recognition These include methods that recognize images based on concise positions in the image as well as their relationship to each other. They are much more resistant to geometric distortion and rotation than robust hash methods. As long as a sufficient number of concise positions in an image are preserved, sections can also be recognized. On the other hand, memory requirements and computational effort are significantly higher than for robust hash methods. For example, they are resistant to changes.

Image classification Here, machine learning, now widely used deep learning, is used to relate higher-level terms to example images, and thus later link new images showing similar content to that term as well. If enough images related to a superordinate term are shown as examples during training, the system can recognize objects of that type even if they differ from the examples in some places.

Requirements

Since the application is to be used for example in customs investigation, auctions and trade fairs, a fast result within a few seconds is essential. The speed must not be achieved at the expense of accuracy, but the errors made by the system must be very small. This is because false positives cause unnecessary effort in the investigation work and false negatives would leave non-tradable objects in circulation. The trained model as well as the image matching should be resource efficient and return a result below three seconds on a server using the CPU and without GPU.

Proposed Approach

As a solution, we propose a system that works on the basis of an app in interaction with a server application (see also Figure1. The app is used by a person who is involved in the investigative work and photographs the object to be checked with the app (1). The image material is then transferred to a server (2) to estimate the origin and time period. There it is compared (3) with reference material assessed by experts (4). If there is sufficient similarity, the information about the reference object is transferred, ideally with an image of the reference object (5). The information obtained in this way allows the user to decide (6) whether the object should be examined more closely and, for example, an expert should be consulted. To implement the classifier, we crawled, cleaned, and processed data before training a deep neural network.

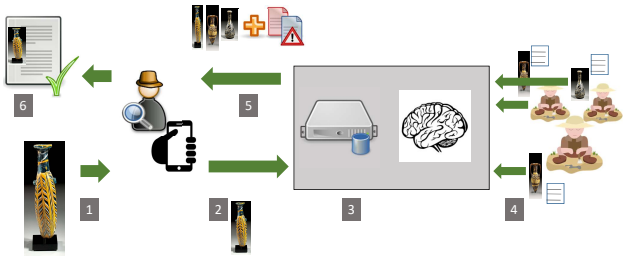

Figure 1. *Flow of data exchange between the user, app and the server.*

Smartphone App

The developed app considers a simple and intuitive user guidance via the main screen 2, where the user can choose from which perspective the following images are captured. A maximum of six images per object can be captured and forwarded to the server for classification. When taking the picture, the user receives information about the contrast and brightness, so that he can take the best possible pictures. This is to ensure that better results are returned from the server after classification. Once a picture is taken, the user is given the option to crop the image so that a potentially misleading background is eliminated from the images sent to the server.

Figure 2. *Main screen of the app with the possibility to select the perspective from which the following picture will be taken.*

Data Acquisition

In order to create a data set for training, we used the online collections database *smb-digital* of the *Staatliche Museen zu Berlin* in Berlin, Germany. The database comprises of more than 250,000 archaeological objects, each with an image and metadata, e.g. a description of features, origin, or age. Because there was no API available for downloading items from the database, we had to build a web scraper to collect the data directly from the website of smb-digital (`http://www.smb-digital.de/eMuseumPlus?lang=en`).

We wrote the scraper in Python, mainly using the *requests* and *lxml* packages for sending web requests and parsing html, respectively. While testing our scraper we noticed a significant delay between web requests that we could also reproduce manually using a browser. When we tried scraping using multiple web requests simultaneously, we saw no increase of the delay per worker. Therefore, we decided to built the scraper with multithreading support to distribute the total number of required web requests among multiple workers running in parallel. This allowed us to reduce the amount of time required to download all items from 70 hours down to only 7 hours using 10 workers. In total, we were able to download the images and metadata of 253,113 objects, resulting in 54 GB of image data in JPEG format and 1.4 GB of uncompressed metadata in JSON format.

Data Preparation

The crawled set contained data that was not ready to be used for training. Almost all images had text labels at the bottom of the image. Cropping the bottom edge fixed the problem without affecting the objects on the images. Some images were in the dataset as placeholders. These were identifiable by the label "no figure available" and were removed from the dataset. A big challenge was the very heterogeneous time information about the origin of the object. Most of the data was not in computer readable format like ("XXXX - YYYY", "XXXX/YYYY", "XX - YY Jhr. / Jahrhundert", era string "Ptolemäerzeit") and not a specific year but rather a range of time. We defined the format for specifying the year as follows: (object time begin, obj time end) and tried to convert the data accordingly into this format. Most strings could be parsed into this tuple by using a regex based approach. , If one specific year was specified in the data, the year was used for object time begin and object time end. For cases where only a name of the era was given, a lookup is needed to translate the era string to a numeric range. We try to get the wikipedia articles for parsing the year range from there. However, the time attributes were still very heterogeneous. It is not possible to classify every possible value. Therefore, the Jenks Natural Breaks Optimization clustering algorithm was used to quantify the continuous time variable into discrete classes using year ranges.

Classifier Architecture

Due to limited training data we decided to take a transfer learning approach. A pretrained model that was trained on a multitude of general object types should have learned universal visual feature representations. Ideally, finetuning with our dataset should only slightly change the convolutional features, mostly in later layers representing higher-level features and thus boost the performance versus a network trained from scratch. We chose the pretrained ResNet [3] network called *BigTransfer* [1] which has been pretrained on the *ImageNet-21k* data set [2]. For the model head, the standard BigTransfer model uses group normalization [4], ReLu activation, global average pooling [7] and a dense layer for the softmax classification outputs. we used the model variants ResNet50 called BiT-M-R50x1 and ResNet152 called BiT-M-R152x2 for training.

In the final paper we will evaluate extension to the above model architecture which previously performed better on a data set with different (worse?) object time binning into classes then the default BigTransfer model. For this experiment all intermediate feature maps of the four ResNet blocks are not only propagated successively to the following block but also combined further. After the model's body, all features are both average and maximum pooled to a common resolution of 120 by 120, concatenated and fed to the network head. The head is extended to apply group normalization, a fully-connected layer, ReLU activation and a fully-connected layer for classification output sequentially. Also, bigger ResNet variants are evaluated.

Training

First, we split the data set in the ratio 80-10-10 for training, validation and testing respectively. The model is optimized for the training set and validation is performed after each training epoch. A new checkpoint is saved if the validation loss has reached a new minimum value. This way we prevent overfitting by not saving models with increasing validation losses while training loss can still decrease. After performing experiments with different hyperparameters and model variants, the test set is used for a final evaluation to estimate the models generalization capabilities and to prevent overfitting the hyperparameters to the validation set. For most hyperparameters we used the standard BigTransfer training scheme: the model was optimized using SGD with momentum of 0.9, mixup [5] was used for regularization by combining data points with a batch size of 512 on a single GPU. Input images are resized to 448x448 followed by a random crop of size 384x384 and horizontal flips. Learning rate is scheduled to first warmup for 500 steps until reaching the base learning rate and then decayed by factor 10 after 3000, 6000 and 9000 training steps. In addition to the BigTransfer paper, we aim to find a good base learning rate by performing a small training run where the learning rate is increased after each step. The resulting plot of the learning rate and loss gives insight for choosing an optimal learning rate [6].

Both model variants are finetuned for 20 epochs. The ResNet50 model's loss was lowest after 4 epochs after which it continued to get bigger. For the ResNet152 model, this point was reached after 4 epochs.

Image Matching

To find the best matched objects, image matching is performed in the results of classification using the features extracted from the trained model by transfer learning. The best top k classes of the classification results are used to form a subset in which image matching is then done to find the most similar objects. In this work, different features are extracted from the trained classification model and evaluated for the image matching task, including the features extracted only from the high layer of ResNet50 and the fusion features extracted from the low and high layers.

The high layer features are obtained from the conv5_x layer

of ResNet50, which gives good semantic representation but lacks the capability of distinguishing objects of the same type. The fusion feature combines the weighted output of the low and high layers, i.e. from conv2_x to conv5_x of ResNet50, in order to better capture the subtle differences in similar objects. Moreover, the extracted features are further downscaled for efficient matching by adding additional average pooling layers.

Combine the individual probabilities to a global probability

We get for each of the maximum six submitted images from the app to the sever a mapping to some collections by the classifier and then for each captured image the best matching images from these collections are searched and sent to the app for display. These images are assigned a probability by the image matching. We want to provide the user with only one probability value per image, from which they can quickly infer the origin and temporal association for the photographed object.

For this purpose, an image from a collection can be recognized several times as a good match and this should be positively taken into account by the system. The calculation to a probability value to a global value is done as follows: $P_{IID}(i) = P_I(i) * P_C(i)$, where P_I is the probability for a specific image returned by the image matching algorithm and P_C is the probability of the collection.

If an image from the dataset is recognized with a probability value greater than 0.2 for multiple perspectives, this should have a more positive effect on the overall result, resulting in a higher probability being assigned to the image: $PG_{IID} = \sum_i P_C(i) * P_I(i)$.

Evaluation

The evaluation include first results of the tests with the system. In the final paper we will give a status update from the 6-month testing phase starting in October with 15 involved Smartphones. The Smartphones are used in e.g. museum where the users have access to the correct label of the captured object. Besides that we run the classification on the test data and conclude the results.

Accuracy

The results of classifying a collection and object time interval on the test data provide 98.94% for the top 5 collections and 80.76% for the top 1 collection using ResNet152x2, 98.60% for the top 5 collections and 78.25% for the top 1 collection using the smaller ResNet50x1 network. Further optimizations can be achieved by changing the model architecture as described and we would discuss these in the full paper.

Performance

The speed of classification and assignment to a collection is done within 100-200ms per uploaded image. The image matching algorithm needs 40ms for the comparison with 1000 images. Depending on the number of images in a collection (average is 1120 images per collection) the algorithm needs on average about 200ms for 5 collections per image. For uploading and downloading further time is required, which is very difficult to estimate, as it depends on the quality of the Internet connection. But with the mentioned speed for the two involved algorithms the performance is very good for the application.

Conclusion and future work

In this paper we have described a system consisting of a smartphone app and a server with the goal of being able to estimate the region of origin and time period. These two parameters are important to estimate the tradability of an object. In the development of the app, attention was paid to a simple and clear user interface. In the development of the classification and image matching, a new approach had to be chosen, since the question cannot be answered with existing pre-trained network.

Instead of using an algorithm to interpret the results, we decided to use an image matching procedure. Thus, the most similar images from the dataset are displayed. The images are found using the features exported from the classifier. The first results on the test data are very positive. Until the start of the test phase, the algorithms will be further optimized and tested. We plan to evaluate and discuss the final architecture and parameters after the evaluation phase.

Acknowledgments

The Federal Government Commissioner for Culture and the Media is funding the project with up to 500,000 euros from funds of the Federal Government's national AI strategy.

References

[1] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, Neil Houlsby, Big Transfer (BiT): General Visual Representation Learning, arxiv, 1912.11370, 2020.

[2] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, ImageNet: A large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition, arxiv, 1512.03385, 2015.

[4] Yuxin Wu, Kaiming He, Group Normalization, arxiv, 1803.08494, 2018.

[5] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz, mixup: Beyond Empirical Risk Minimization, arxiv, 1710.09412, 2018.

[6] Leslie N. Smith in Cyclical Learning Rates for Training Neural Networks, arxiv, 1506.01186, 2017.

[7] Min Lin, Qiang Chen, Shuicheng Yan, Network In Network, arxiv, 1312.4400, 2014.

[8] Vanessa, Hanson, Looted Antiquities: Economic Opportunity for Terrorists, Scholarly and Creative Works Conference 2020.

[9] L. A. Amineddoleh, Cultural heritage vandalism and looting: the role of terrorist organizations, public institutions and private collectors. Santander Art and Culture Law Review, 1(2), 27-62, 2015.

[10] N. Brodie, I. Sabrine, The illegal excavation and trade of Syrian cultural objects: a view from the ground. Journal of Field Archaeology, 43(1), 74-84, 2018.