Pattern Recognition Letters 135 (2020) 337–345

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Markerless detection of ancient rock carvings in the wild: rock art in Vathy, Astypalaia

Giorgos Tsigkas[a], Giorgos Sfikas[a,c,*], Anastasios Pasialis[b], Andreas Vlachopoulos[b], Christophoros Nikou[a]

[a] Dpt. of Computer Science and Engineering, University of Ioannina, 45110 Ioannina, Greece
[b] Dpt. of History and Archaeology, University of Ioannina, 45110 Ioannina, Greece
[c] Information Technologies Institute, CERTH, 57001 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 14 July 2019
Revised 9 February 2020
Accepted 25 March 2020
Available online 23 April 2020

MSC:
41A05
41A10
65D05
65D17

Keywords:
Markerless object detection
Ancient rock carvings
Feature point-based matching

ABSTRACT

In this paper, we discuss the problem of object detection in a cultural heritage application context. In particular, the objects to be detected are ancient rock carvings, discovered at the archaeological site of Vathy, Astypalaia in Greece. Without the help of a marker or a human expert, the rock carvings are extremely difficult for a visitor of the site to discern from their surroundings. We explore the possibility of using a computational method that could replace the human expert and detect the rock carvings of interest without the aid of a specific marker. We present a dataset of images that is comprised of annotated photographs of the rock carvings, taken *in situ* and under differing poses and lighting parameters. Two methods for detection are applied; the first method makes use of a supervised, deep learning-based model, while the other relies on feature point-based matching to an annotated template, in the context of which we propose a simple image matching distance. We show that each method is applicable under different conditions, and evaluate their effectiveness with numerical trials.

1. Introduction

In the recent years, computational methods have been proposed and used in the context of cultural heritage in a diverse spectrum of topics and applications [2,4,6,8,10,12,21]. Various methods, algorithms and associated software exists today to help enhance the experience of tourists to a museum, town or other culturally significant site [6], or aid cultural heritage professionals in documenting, cataloguing or preserving cultural heritage objects [19]. 3D reconstruction for digital preservation [8] is one such prominent application, typically applied to sculptures, pottery, and other man-made artifacts. In the context of preservation, 3D reconstruction has also been used as part of a more extended pipeline, providing models for further processing at a later stage; for example, in [3,4], prehistoric fresco shards are scanned and used as part of a model that takes into account the affinity and ease of matching between fresco shard pairs, to the end of reconstructing the whole fresco from its parts; in [2], ancient inscriptions are processed to the end of further statistical analysis of the reconstructed surface. Techniques of object retrieval have been applied on cultural heritage objects such as pottery artworks [17], potentially easing searches in an extended artifact collection. Significant research targets digitized corpora of manuscripts, with computational tools having being elaborated for digitized document analysis [7,16].

In the context of cultural heritage as preserved and exhibited in museums, new technology is also being increasingly used [18]. Vision-based applications are in the forefront here [27], with computer vision techniques being exploited in order to enchance visitor experience. Augmented reality methods are based on cues from the user hardware, including visual inputs that may rely on the placement of specific markers. Such markers may be a very effective way to determine a static position from the camera [26], which is by definition easily discernible from their environment. However, while markers may facilitate vision-based detection, it is not always convenient to attach a marker everywhere it is required [27]. In this sense markerless vision systems are advantageous, with the caveat that the detection component of the system may now correspond to a non-trivial problem, depending largely on the context and nature of the application.

In this paper, our focus is on object detection in the context of cultural heritage. We assume that no markers are used to pinpoint

* Corresponding author.
E-mail address: sfikas@cs.uoi.gr (G. Sfikas).

https://doi.org/10.1016/j.patrec.2020.03.026
0167-8655/

the objects of interest, hence detection is to be performed in the standard mode expected in other applications of computer vision, i.e. based on the image content itself. We present a new database of images, created at the archaeological excavation site at Vathy, Astypalaia, Greece, which includes images of rocks with rock carvings, acquired *in situ* and under a variety of poses and lighting conditions. While our objects of interest, i.e. the ancient rock carvings should be the main point of interest for a potential visitor of the site, these are next to impossible to discern withouth the help of an expert. This is due to milennia of exposure that have weathered the rock, making the carving difficult to see even at a very close distance. A succesful object detection system would hence be very useful in the respect of being in perspective used in the context of an AR application, for example embedded to a smartphone app [5]. For example, a tourist visiting the archeological site in the absence of an expert's guidance would simply have to hold his or her smartphone with its camera facing at objects of potential archaeological interest. If one of the rocks is found to contain a carving, the application would alert the user by showing a bounding box around the rock as well as a short description and other information about the carving. In this manner, the tourist experience on the site can be significantly more autonomous compared to as being dependent to a tour guide. Furthermore, no visible markers are required to be added anywhere on the site; the presence of markers, such as signposts, can be aesthetically displeasing both for site visitors and with respect to the site itself.

Let us also note that the problem is made even more challenging due to the lanscape itself; it is made up for the most part of large rocks of grey dolomite limestone, uncarved but similar to the ones we aim to detect. Furthermore, the most abundant objects beside rocks, are low cedar bushes, that naturally look very similar to one another. These factors, along with the shallow carving or pecking of the rock engravings, constitute a challenging detection problem.

In order to solve the stated problem, we propose the use of two different vision systems, one relying on a deep learning-based approach, while the other being based on more standard computer vision tools. We use the YOLO model [13,14] as our deep learning-based approach. YOLO is a state-of-the-art object detector recently proposed, and already used in a very wide range of applications. While deep neural network-based models are indeed very succesful in solving many different tasks [9], they are notorious for requiring very large datasets for training. On the other hand, training a neural network is equivalent to solving a difficult optimization problem in a very high-dimensional parameter space, meaning in practice that training is not always a straightforward task, depending on numerous hyperparameters, good initialization, and training strategy heuristics. For these reasons, we have also used a non-deep learning based method. The core of the second method is Scale Invariant Feature Transform (SIFT) feature extraction and matching with an annotated template image using Random Sample Consensus (RANSAC) [11]. On the context of this latter method, we propose a simple distance metric in order to determine the most appropriate annotated template. We validate the usefulness of this metric by computing manifold embeddings with respect to the metric using Isomap [20], as well as with numerical results. As we shall discuss in more detail in the following sections, each of the two proposed detection methodologies comes with its own merits and drawbacks according to the assumptions we make about the available images.

The paper is organized as follows. In Section 2, the dataset acquisition process is presented. In Section 3, we discuss the methods that we have used for markerless detection of the objects of interest, and in Section 4 we present and compare the methods with numerical and qualitative results. We close the paper with Section 5, where we discuss conclusions and future work.

2. Dataset

The new dataset consists of images captured at the Vathy excavation site, located on the island of Astypalaia, Greece. Archaeological research has commenced in Vathy in 2011, around the ruins of an hellenistic tower. Since then, research has uncovered findings from numerous historical eras, including most notably an Early Cycladic fortified settlement [23], as well as later, albeit less numerous findings. One of the most important findings of the fieldwork at Vathy was the identifying and studying of a large number of ancient rock carvings. Most of the carvings have been identified as prehistoric, two of them being inscriptions dated to the archaic and classical periods (early 6th - early 4th c. BCE) [25]. In the context of the proposed image dataset, we are interested in specifically three rocks and the corresponding rock carvings; two of them are dated to the Early Bronze Age (3rd mill. BCE), and one inscription dated to the classical era (early 4th c. BCE). The three rock carvings can be examined at Fig. 2. In the remainder of this section, we shall discuss in more detail the archaeological context corresponding to the rock carvings of interest, as well as describe the process of creating the dataset.[1]

2.1. About the excavation site

Vathy is a naturally protected peninsula controlling the narrow access from the open sea to the homonymous gulf on the northwest rocky coast of Astypalaia, thus ensuring full monitoring of a wide area of sea and land (Fig. 1). At the easternmost tip of the Pyrgos promontory, on Cape Elliniko, an acropolis was founded in the 3rd millennium BCE, the boulder-built circuit walls and megalithic retaining walls of which are visible today. On the upper level of the headland, a tower with surrounding ancillary facilities was erected in the late 4th c. BCE. The intensive surface survey which has been carried out at Vathy since 2012 and the systematic excavation since 2014 have revealed numerous monuments and finds, which further establish the Early Cycladic elements, but also point to influences from other areas of the Aegean [22,23,28].

The recovery of Early Cycladic marble figurines and numerous rock engravings of spirals, oared ships, daggers, arrows, axes and other motifs over a wide area of dolomitic limestone on the cape, both quarried, built and in natural state identify Vathy as a significant site of the Early Bronze Age Aegean, with intense Cycladic features [24]. Archaic (6th c. BCE) and classical (4th c. BCE) inscriptions were also located on the peninsula indicating that human activity continued in later centuries (Fig. 2). The 4th-c. BCE inscription of someone named Dion, probably a guard patrolling the strategic point of the peninsula [25] and the emblematic rock art representations of a long oared ship and of the daggers, that decorate two monumental gateways of the 3rd mill. BCE acropolis, have been chosen as the case study motifs for the purposes of the paper in hand.

2.2. Procedure of dataset creation

The proposed dataset is comprised of images shot at the Vathy excavation site. In total, we include 1078 images in the dataset, depicting in their majority one of the three rock carvings that were discovered on the site as a result of archaeological fieldwork. These three rock carvings will be named, in the context of this work, "Dion", "Ship", and "Dagger". These names are conventional, and are loosely based on the content of each of the rock carvings. They are each located on top or the side of a different rock, and each of

[1] The full dataset along with the prescribed partitions can be downloaded at http://cs.uoi.gr/~sfikas/astypalaia.

Fig. 1. Location of the excavation site where the dataset was collected. The site is located at the north part of the island of Astypalaia,Greece (36°37'03"N,26°23'43"E). (a) The box delineates the peninsula where the excavation site is located ("Cape Elliniko") (b) The peninsula seen (marked with a box in (a)) from the opposite coast.

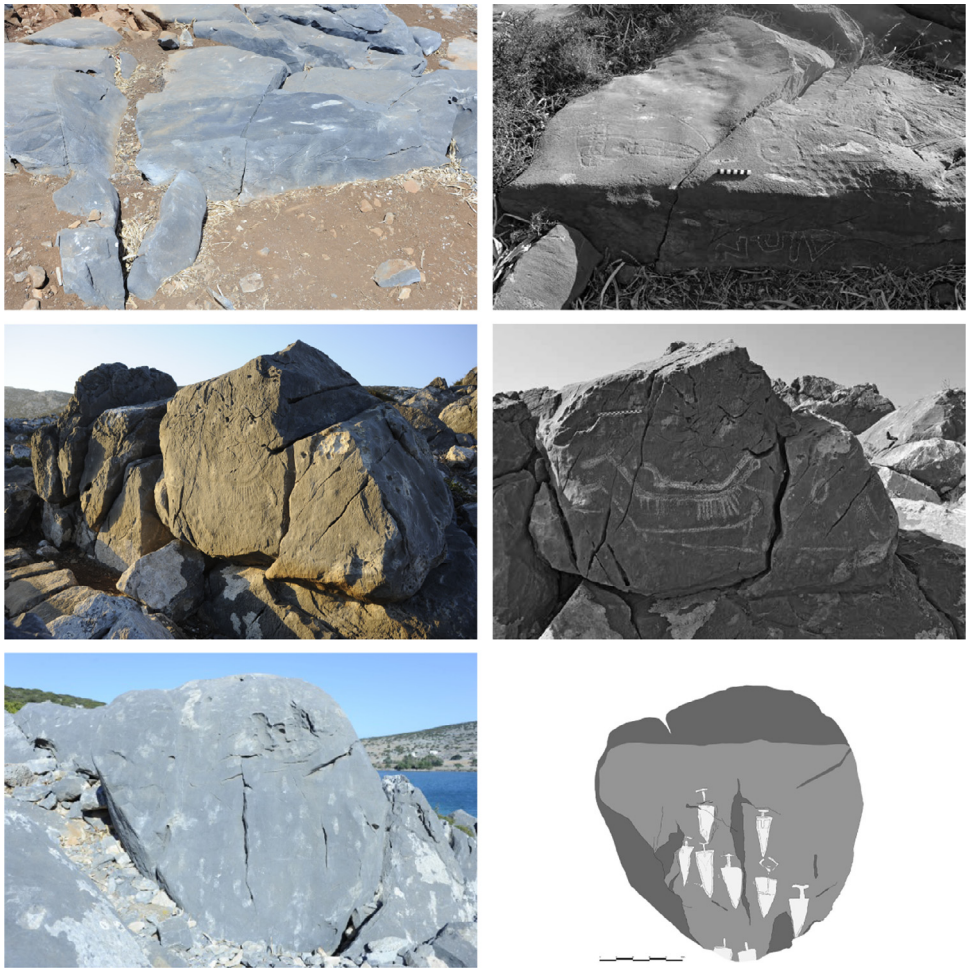

Fig. 2. The three rock carvings of interest. On rows from top to bottom we see the carving of "Dion", "Ship" and the "Dagger". The pictures on the left column are actual images included in the dataset. The pictures on the right column (originally in [25] and [22]) depict the same rocks and rock carvings either marked by the excavation workers and photographed under conditions that would make the carvings more easily visible, specifically for the picture (two leftmost images) or depicted as a sketch (rightmost image). The unmarked and marked images are shown here in juxtaposition so that the reader can better appreciate that in all cases the actual rock carvings are extremely hard to point out.

the rocks is located at a distance of several tens of meters from one another. The rock carvings of interest can be examined in Fig. 2. Note that all three rock carvings are extremely hard to discern by a non-expert visitor of the site, even under good lighting conditions and/or at a very close distance.

In terms of hardware, a *NIKON D700* camera (focal length 5 mm) was used to capture the dataset images. All images were captured at a resolution of 2128 × 1416 pixels. Images were captured during the month of July, 2016, and they are grouped – aside from the depicted rock carving – according to two other varying parameters: time of capture, and distance from rock carving of interest. We took photos in four distinct times during the day: these are "daybreak" (around 6am), "morning" (around 10am), "noon" (around 2 p.m.) and "afternoon" (around 6pm). On a varying degree, rock carvings may be easier or harder to see during different times in a day, and this is also a function of the rock carving location and orientation. The next major parameter that varies for pictures of this dataset is the distance of the photographer from the

Table 1
The number of images depending on the rock carving they depict, the time they were captured and the distance between the camera and the rock. "c", "m" and "f" denote distance of camera from the photographed rock, corresponding to "close", "mid-range", "far" respectively (see text for details).

	Daybreak			Morning			Noon			Afternoon		
	c	m	f	c	m	f	c	m	f	c	m	f
Dion	8	26	38	35	26	24	24	15	16	9	16	17
Ship	16	23	22	14	26	28	12	19	16	11	21	23
Dagger	14	17	29	16	34	29	14	16	21	18	19	17

rock carving of interest. We used three distinct distance groups, named: "close" (1 – 2 m), "mid-range" (4 – 5 m), "far" (10 – 15 m). Having included a variety of times/lighting conditions, as well as distances to the object of interest, we help to avoid a positive or negative bias in terms of automatic detectability. Furthermore, dataset variability can help in terms of emulating conditions for future images that may be used as input to a computational pipeline, trained or finetuned with respect to the supplied images (Table 1).

3. Proposed methods

We propose two methods to solve the problem of automatic detection of rock carvings. The first detection method is based on a deep neural network architecture, that has given state-of-the-art performance in object detection tasks in the literature [14]. As all neural network models, this method however requires a large set of images that are beforehand annotated with the true rock carving locations, used as a training set. In order to simulate a setting where a large training set is unavailable, we propose an alternative, second method. SIFT feature points [11] are computed on a test image, as well as an annotated template; RANSAC is then used to match the two images [11] and to finally produce the predicted object location. In what follows, we discuss the two methods in detail.

3.1. Deep learning-based object detection

As our deep learning-based object detector, we use YOLOv2 and a light-weight version of the same model, TinyYOLO-v2[2] [14]. In this paper, we shall refer to the two models as YOLO and TinyYOLO for brevity. YOLO is a state-of-the-art real-time object detection system based on a convolutional neural network. It is also a highly generalizable system, which means it is less possible to make detection errors in different domains from what it was trained on. Moreover, YOLO attains fewer false positives in the background of the image because it performs the prediction accessing the full input image, and not selected areas like other detectors. Hence, we believe that selecting YOLO as a representative supervised detector for the purposes of the current work is a sensible and quite straightforward choice. YOLO is fed with the image where potential objects of interest exist, and returns as output possible positions, geometry and predicted object classes. The neural network (called Darknet-19) is comprised of a total of 19 convolutional layers, topped with batch-normalization layers. These are further coupled with 5 max-pooling layers, while dropout layers are also used to improve model generalization.

Training is performed in each case using the training subsets of the proposed partitions (see Section 4.1). Part of the training set is employed as a validation set, used as a training-time benchmark to either reduce the learning rate or terminate training. For TinyYOLO, RMSprop was used for training, with initial learning rate set at 10^{-3}. Learning rate was then gradually reduced to 10^{-4} and finally 10^{-5}, after which training terminates when validation set detection rates are not further improved. For YOLO, an analogous training schedule was used[3]. Network parameters were initialized with weights that were pre-trained on the COCO dataset [14].

3.2. Feature point-based detection

A number of annotated images is here also necessary. In this method, these images are however used in a different manner. Instead of being used to drive an optimization scheme, they are used as template images. This is to be understood in the sense of images that should be matched by a specific transformation with respect to the input (test) image. Once the required transformation is computed, the ground truth bounding box that is found on the template is also transformed to align with the test image, thus becoming the predicted region. Examples of matching against an annotated template using this scheme can be examined in Fig. 4.

The first step of the method requires SIFT keypoints to be computed on the test image, as well as on all candidate matching templates. The matching candidates are all pre-annotated images (i.e. the "training" set). We have used the OpenCV library[4] to compute SIFT keypoints. We compute at the most, 5000 keypoints per image. In order to speed-up the algorithm, images of half the original width and height are used. In order to select the most appropriate template, we compute a similarity measure between the test image versus all candidate templates. This similarity measure is defined as the number of keypoints that match between images. Matches are computed using the RANSAC algorithm, which estimates a projective transformation between the two-point sets by following the next steps.

(1) We choose 3 SIFT keypoints randomly from the first image and then, from the second image, we choose the 3 best matching points to the previous 3 points. The matching factors are the features of the keypoints and the Euclidean distance between the 3 points. Subsequently, we add the best-matching pairs of points to a "consensus" set. (2) Based on the consensus set, we compute the parameters of the projective transformation T. (3) We apply the T transformation to all the points of the first image and we add the pairs of points, to the consensus set, that match the best between the transformed points and the points of the second image. (4) We return to step 2, until T transformation's paramaters will not change at the next iteration. A number of matching keypoints below 10 is considered too low, at which case the two images are rejected as non-matching. Our premise is that more matching keypoints should indicate a better image (content) affinity between images, resulting from a similar camera position and orientation for the two images. In order to test this premise, we have computed two-dimensional embeddings for our images using Isomap

[2] We used the popular implementation of https://github.com/thtrieu/darkflow.

[3] Stochastic Gradient Descent (SGD) was used, as we found that the implementation of RMSprop for YOLO particularly was memory-inefficient.

[4] https://opencv.org/.

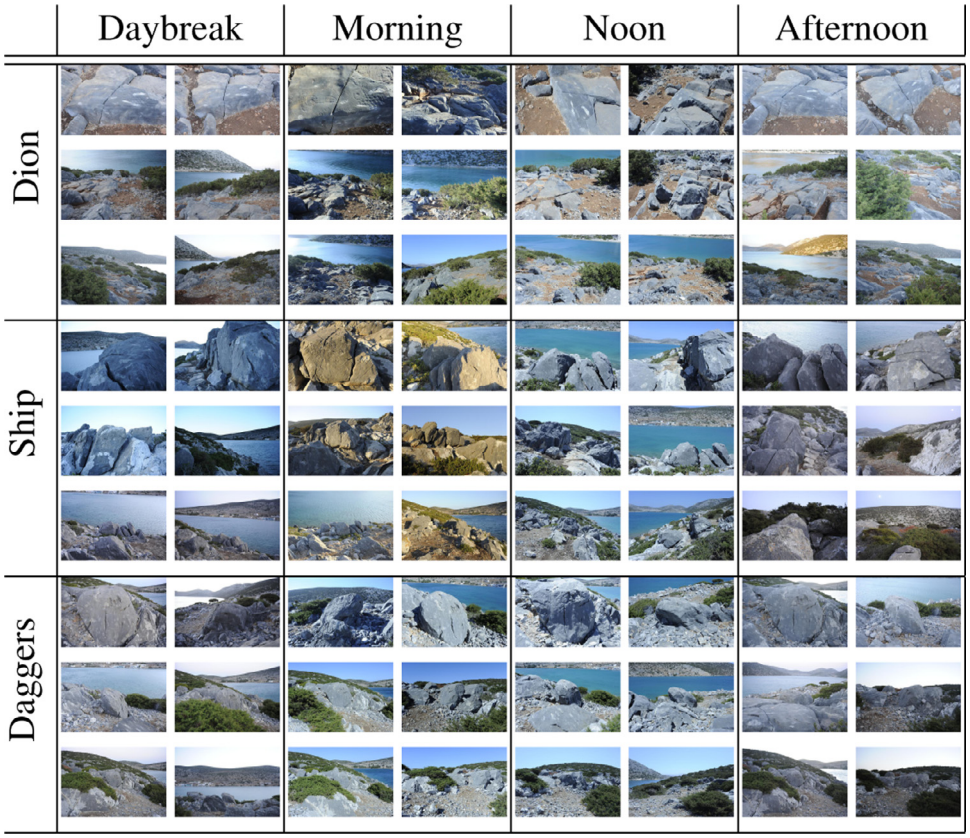

Fig. 3. Dataset sample images.

[20]. Isomap estimates a non-linear manifold, assigning each image to a single point in a low-dimensional space (here, $\in \mathbb{R}^2$). The estimated low-dimensional coordinates are constrained to minimize discrepancy between inter-point distances that are provided as input to the algorithm, and distances on the calculated low-dimensional space. The embedding can be computed through the following steps. First, we construct a neighborhood graph by connecting pairs of points, if the first point is one of the 'k'[5] nearest neighbors of the second point. Subsequently, we compute the shortest path distances between all the points in the graph and we put them on a matrix D_G which is now called a distance matrix. At the last step, we construct the requested d-dimensional embedding by applying the classical MDS algorithm to the distance matrix D_G. In Fig. 5, Isomap embeddings computed over a sample of the image dataset are shown, juxtaposed on the map of the excavation site. As Isomap requires distances between points as its input, we have used a simple exponential transformation $d(I_1, I_2) = e^{-\lambda m}$ to compute a distance between each image pair I_1, I_2. λ is a transformation parameter (here set to $\lambda = 0.05$) and m corresponds to the number of matches computed using the previous scheme. Isomap requires also its free parameter, 'k', to be selected properly. Using an automatic method for selecting the optimal value for this parameter [15], the resulting parameter for the "Dion" carving embedding is 4, for the "Dagger" carving embedding it is 3 and finally for the "Ship" carving embedding it is 2. Note that the embedding points form approximate circles, with images taken from close positions appearing as nearby points on the embedding. The embedding relative positions appear to be correlated to the real geographical positions from which the photographs were acquired, as they were photographed as shots of a constant radius around each rock carving of interest. Hence, this result validates the usefulness of the proposed distance measure.

Note that in numerous cases, the object of interest by itself (the rock carving) is not useful as a region of features to be matched, that is useful in terms of using its image content to detect it by computational means. This is due to the fact that we aim to detect rock carvings, which may be difficult to see or even invisible to the naked eye (especially the "dagger" carving, Fig. 2). For this reason, the image *context* of the rock carving, rather than the rock carving itself, is more useful to detect it. Indeed, keypoints of the rock carving surroundings may be easier to match. For example in Fig. 4, last image pair, SIFT keypoints for a nearby rock are correctly matched, while the rock that includes the carving is not useful in terms of SIFT matching. This is inconsequential to the final output, as matching the context around the rock carving is fortunately enough in many cases; in the aforementioned example, the correct[6] geometric transformation is computed, and used to transform the template bounding box correctly. On the other hand, matching keypoints that correspond to objects that are too far away from the rock carvings is detrimentary to computing a good geometric transform. Such objects are usually details on "background" hills or coastlines. This is not at all suprising, as image elements that are on a significantly different distance w.r.t. to the foreground will correspond to a much different (smaller, if these objects are on the background) apparent movement. In turn, this will lead to an erroneous estimated transform. For this reason, we aim to reject key-

[5] 'k' is the only free parameter of Isomap and it is the number of the nearest neighbors of each point.

[6] Correct in the sense of acceptable in practice; in theory, the camera movement that describes the difference between the two images may not correspond to a projective transformation.

Fig. 4. Detection results using the feature point-based method. An annotated template image is necessary (left column) to perform detection on an input test image (right column). White bounding boxes are manually annotated ground truth rock carving positions. Black quadrilaterals on test images (right column) correspond to the transformed box resulting from the transformation of the ground truth bounding boxes available on the template images (left column). The detected SIFT keypoints that are part of the concensus set are shown, along with the RANSAC-based matches (lines connecting points from one image to the other) between keypoints on each template and test image pair.

points that correspond to background, distant image elements, by using only keypoints found in the lower 60% of the image height.

4. Numerical evaluation

4.1. Dataset partitions

In total, there are 1078 images in the dataset. Samples of the images can be examined in Fig. 3. All dataset images have been manually annotated using the LabelImg[7] annotation tool. The provided annotations are in the form of a bounding box around the object of interest, as well as its class, i.e. the specific rock carving being photographed.

For our experiments, we have used two different partitions of this dataset into varying-sized training and test sets. The two partitions differ mainly with respect to their size, hence named *Full* and *Minimal* partition respectively. The *Full* partition contains all 1078 images. We have chosen images so as to ensure that on both the training and the test partitions all landmarks, all times and all distances are represented. The *Full* partition has thus been split into a 90% to 10% ratio between the proposed training and test sets, with 967 images and 111 images to each subset respectively. The *Minimal* partition includes only a fraction of the full set, with a total of 111 images. These are partitioned again at a similar ratio between a training and a test set, with 9% of the full set used as the training set, and about 1% used as a test set. These percentages correspond to 99 and 12 images for the two subsets respectively.

A version of the full dataset has also been created, where the training set is augmented with artificially created images. Data augmentation is standard practice in contexts where a model requires large training sets, such as is the case with neural networks [9]. For each image in the training set of the *Full* partition, three new images randomly rotated were created, at a maximum absolute rotation angle of 30 degrees. Also, three new images from randomly cropped areas of the initial image and then upsampled at the resolution of the initial image were created. We refer to this partition, totalling 6214 images, as *Full-Aug* in our experiments.

4.2. Experiments

About the training and validation process of YOLO and TinyYolo detectors, we decided to train the deep networks dynamically by checking on the validation errors and not with some predifined value of epochs. This is the reason that the numbers of epochs shown in Table 3 vary. More specifically, we started the training process with a learning rate set to 10^{-3}. During training, if the val-

[7] https://github.com/tzutalin/labelImg.

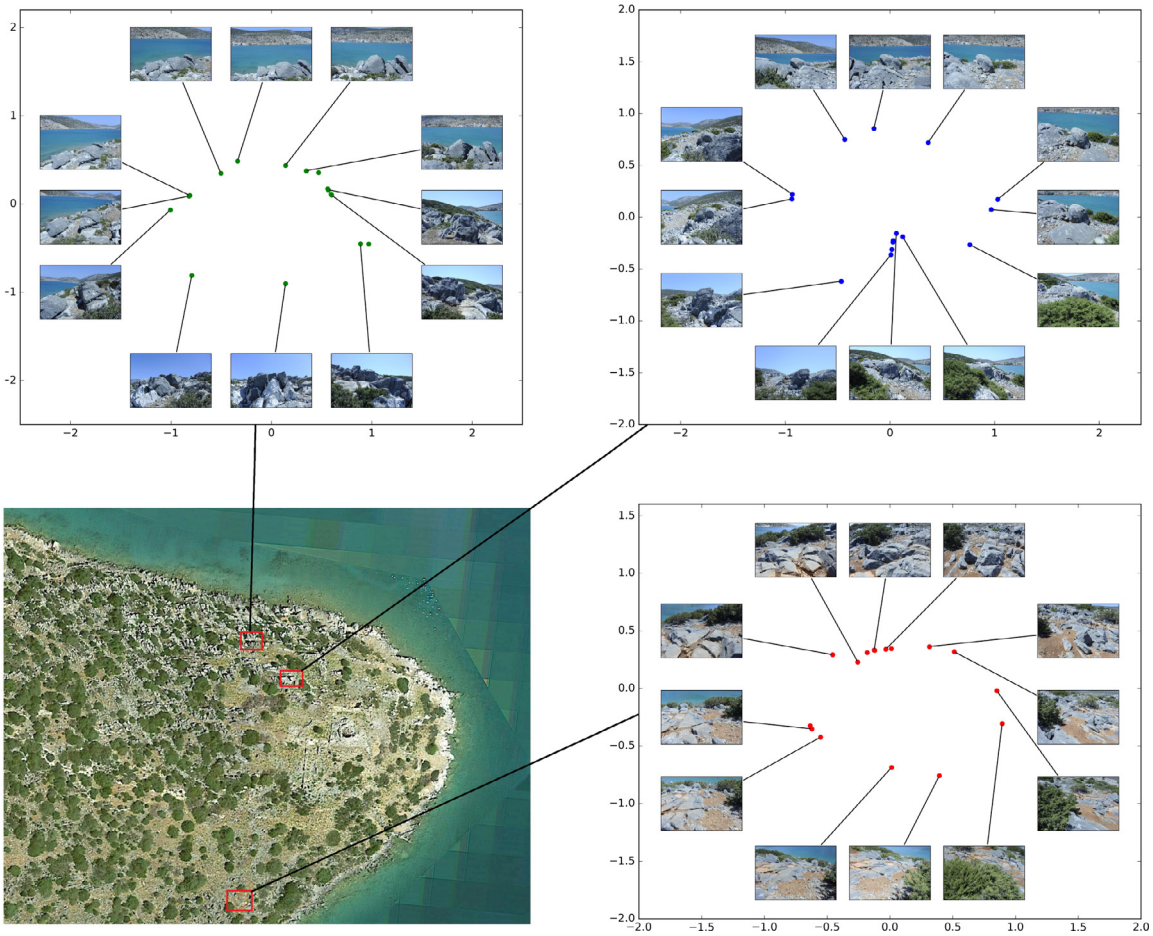

Fig. 5. Two-dimensional embeddings for a sample of the dataset images, using Isomap and the proposed SIFT-based distance measure as input. The position of the rock carvings on the site map is shown (bottom left quadrant), to each of which one embedding is computed. Images that were acquired from positions close to one another are automatically assigned to close positions to their respective embeddings, thus correlating to the original camera geographical positions. The Dion's embedding is the one with the red colored points, the Daggers's embedding is the one with the blue points and finally the Ship's embedding is the one with the green points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

idation error increased but the training error decreased, we would stop this process, then reduce the learning rate to 10^{-4}, and eventually recommence from the training step right before results start to degrade. This process would be repeated until we stopped the training while having set the learning rate at 10^{-5}. Total training time for these detectors are listed in Table 4.

We have evaluated numerically the proposed baseline methods. Detection quality results, measured in terms of Intersection over Union (IoU) scores, can be examined at Table 2. Furthermore, we count a non-zero intersection value only when also the detected object *class* is also correct, i.e. which of the three possible rock carving is detected. Images containing no rock where nothing has been detected, i.e. True Negatives, by convention are set to correspond to $IoU = 1.0$. F-measure scores are also computed [11], over Precision and Recall values averaged over a set of possible IoU values (from IoU = 0.5 to 0.95 in steps of 0.05).

We have run tests comparing all methods on the *Minimal* partition. Mean and standard deviation of IoU scores are shown. Clearly, the feature point-based method gives far more accurate results than both deep learning-based models. This can be attributed to the small size of the *Minimal* partition, with its training set of approximately 100 images clearly not enough to train well a deep network. The situation is different when we need to run tests on the bigger partitions. While a learning-based matcher exploits the size of the bigger training set to lead to better results, the performance of the feature based-matcher does not improve, but rather it worsens. Furthermore, it certainly becomes much slower when run on a bigger set, as more candidate templates need to be considered and matched. YOLO and Tiny YOLO have comparable performance, with similar (bad) performance on the *Minimal* set, and with Tiny YOLO somewhat outperforming YOLO on the *Full* set. YOLO however seems to outperform Tiny YOLO on the largest *Full-Aug* partition, perhaps due to the greater learning capacity of YOLO. In Fig. 6 an example of YOLO detection trained on the *Full-Aug* partition can be examined.

Table 2
Numerical comparison between different methods (rows) and used partitions (columns), in terms of the IoU and F-measure metrics. SIFT-based refers to the discussed feature point-based method, while the other two are variants of the deep learning-based YOLO model (see text for details).

IoU			
	Minimal	Full	Full+Aug
SIFT-based	**0.65 ± 0.2**	0.38 ± 0.3	-
Tiny YOLO	0.25 ± 0.3	**0.59 ± 0.41**	0.58 ± 0.4
YOLO	0.29 ± 0.35	0.56 ± 0.4	**0.65 ± 0.38**
F-measure			
	Minimal	Full	Full+Aug
SIFT-based	**0.57 ± 0.24**	0.38 ± 0.2	-
Tiny YOLO	0.17 ± 0.21	**0.42 ± 0.24**	0.36 ± 0.24
YOLO	0.29 ± 0.23	0.36 ± 0.26	**0.48 ± 0.26**

Table 3
Total number of training epochs for Yolo and TinyYolo.

	Minimal	Full	Full-Aug
YOLO	222	101	77
TinyYOLO	10	153	24

Table 4
Total training time for Yolo and TinyYolo in hours.

	Minimal	Full	Full-Aug
YOLO	2.87	10.97	48.92
TinyYOLO	0.31	17.41	16.03

Fig. 6. Detection results using the YOLO detector. The rock carvings depicted from top to bottom are “Dion”, “Daggers” and “Ship”. White bounding boxes are manually annotated ground truth positions, while coloured bounding boxes correspond to YOLO detections.

In terms of test-time processing speed, the YOLO-based networks both outpaced the SIFT-based method. Tested on the same CPU (Intel Core i7 6700K Processor), YOLO and Tiny-YOLO required respectively 0.39 and 0.06 seconds per frame. When run on a standard GPU (Asus GeForce GTX1060 6GB Dual), these times drop to 0.035 and 0.017 seconds per image. The SIFT-based method is considerably slower, requiring SIFT computations and matchings versus all annotated templates. On average this process takes approximately 2 seconds per matched template (on CPU), which amounts to a total of 207 seconds per detection on the *Minimal* partition.

5. Conclusion and future work

We have presented the challenging problem of automatically detecting the ancient rock carvings found at Vathy, Astypalaia, a task that is extremely difficult without the aid of an expert or a marker. A dataset was created to this end, comprising more than 1000 photographs of the rock carvings. The dataset will be made publicly available. Two baseline methods were tested and evaluated numerically with respect to their detection efficiency. The deep learning-based method was found to be more efficient though requiring as many input instances as possible for training. The feature point-based method was however more efficient than the deep learning-based approach on a setting were only a few already annotated images were available. The results of the latter method have shown that the image context around the rock carving is perhaps more useful than the rock carving itself with respect to detection. This conclusion is justified, as the photos included in the dataset are made up of a scenery that mostly includes rocks and bushes, which when seen separately, are practically indistinguishable from one another.

For future work, the challenges that remain aside from improving detection accuracy is making sure that any object detection can in perspective be used as part of a real, AR application that will aid the visitor of the archaeological site. We believe that this entails at least two constraints: ensuring that any detection method is fast aside from being accurate, as well as compact enough to be used by handheld hardware with limited computational resources, such as a smartphone. Also, the current dataset could be used to apply photogrammetry and 3*D* reconstruction techniques [1], to the potential application of Virtual Reality (VR) guided tours.

Declaration of Competing Interest

None.

Acknowledgments

This research has been partially co-financed by the EU and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call OPEN INNOVATION IN CULTURE, project *Bessarion* (T6YBΠ-00214). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan XP GPU used for this research.

References

[1] L. Barazzetti, M. Scaioni, F. Remondino, Orientation and 3D modelling from markerless terrestrial images: combining accuracy with automation, Photogrammetr. Record 25 (132) (2010) 356–381, doi:10.1111/j.1477-9730.2010.00599.x.
[2] A. Barmpoutis, E. Bozia, R.S. Wagman, A novel framework for 3d reconstruction and analysis of ancient inscriptions, Mach. Vis. Appl. 21 (6) (2010) 989–998.
[3] B.J. Brown, C. Toler-Franklin, D. Nehab, M. Burns, D. Dobkin, A. Vlachopoulos, C. Doumas, S. Rusinkiewicz, T. Weyrich, A system for high-volume acquisition and matching of fresco fragments: Reassembling theran wall paintings, in: ACM Transactions on Graphics (TOG), 27, ACM, 2008, p. 84.
[4] A.G. Castañeda, B.J. Brown, S. Rusinkiewicz, T.A. Funkhouser, T. Weyrich, Global consistency in the automatic assembly of fragmented artefacts., in: VAST, 11, 2011, pp. 73–80.

[5] D. Chatzopoulos, C. Bermejo, Z. Huang, P. Hui, Mobile augmented reality survey: from where we are to where we go, IEEE Access 5 (2017) 6917–6950.
[6] F. Colace, M. De Santo, S. Lemma, M. Lombardi, An adaptive app for tourist contents contextualization, in: Proceedings of the Second International Conference on Internet of Things, Data and Cloud Computing, in: ICC '17, ACM, New York, NY, USA, 2017, pp. 80:1–80:10.
[7] C. De Stefano, M. Maniaci, F. Fontanella, A.S. di Freca, Reliable writer identification in medieval manuscripts through page layout features: the "Avila" bible case, Eng. Appl. Artif. Intell. 72 (2018) 99–110.
[8] L. Gomes, O.R.P. Bellon, L. Silva, 3D Reconstruction methods for digital preservation of cultural heritage: a survey, Pattern Recognit. Lett. 50 (2014) 3–14.
[9] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep Learning, MIT press Cambridge, 2016.
[10] G. Guarnera, F. Stanco, D. Tanasi, G. Gallo, Classification of decorative patterns in kamares pottery, in: Proceedings of SCCG 26th Spring Conference on Computer Graphics, 2010.
[11] R. Klette, Concise Computer Vision, Springer, 2014.
[12] G. Pavlidis, A. Koutsoudis, F. Arnaoutoglou, V. Tsioukas, C. Chamzas, Methods for 3D digitization of cultural heritage, J Cult Herit 8 (1) (2007) 93–98.
[13] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.
[14] J. Redmon, A. Farhadi, YOLO9000: better, faster, stronger, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7263–7271.
[15] O. Samko, A.D. Marshall, P.L. Rosin, Selection of the optimal parameter value for the isomap algorithm, Pattern Recogn. Lett. 27 (9) (2006) 968979, doi:10.1016/j.patrec.2005.11.017.
[16] G. Sfikas, A.P. Giotis, G. Louloudis, B. Gatos, Using attributes for Word Spotting and Recognition in polytonic Greek documents, in: 2015 13th International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2015, pp. 686–690.
[17] K. Sfikas, I. Pratikakis, A. Koutsoudis, M. Savelonas, T. Theoharis, Partial matching of 3d cultural heritage objects using panoramic views, Multimed Tools Appl 75 (7) (2016) 3693–3707.
[18] L. Shu, Van Gogh vs. candy crush: How museums are fighting tech with tech to win your eyes, 2016, Accessed 20 Jun 2019.
[19] F. Stanco, S. Battiato, G. Gallo, Digital imaging for cultural heritage preservation: Analysis, restoration, and reconstruction of ancient artworks, CRC Press, 2011.
[20] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[21] M. Terras, Image processing in the digital humanities, Digital Humanities in Practice 4 (2012).
[22] A. Vlachopoulos, in: Excavation at Vathy, Astypalaia ("Ανασκαφή στο Βαθύ Αστυπάλαιας")., Proceedings of the Archaeological Society at Athens, 2016, pp. 327–356.
[23] A. Vlachopoulos, in: Excavation at Vathy, Astypalaia ("Ανασκαφή στο Βαθύ Αστυπάλαιας")., Proceedings of the Archaeological Society at Athens, 2017, pp. 273–300.
[24] A. Vlachopoulos, A. Angelopoulou, Early cycladic figurines from Vathy, Astypalaia, in: M. Marthari, C. Renfrew, M. Boyd (Eds.), Early Cycladic Sculpture in Context from Beyond the Cyclades: Mainland Greece, the North and East Aegean, Papers Presented at a Symposium Held at the Archaeological Society at Athens, Archaeological Society at Athens, 2019.
[25] A. Vlachopoulos, A. Matthaiou, Neotera Archaeologika Astypalaias ("Νεώτερα Αρχαιολογικά Αστυπάλαιας"), HOROS, Athens 25 (2013) 224–252.
[26] X. Zhang, S. Fronz, N. Navab, Visual marker detection and decoding in AR systems: A comparative study, in: Proceedings of the 1st International Symposium on Mixed and Augmented Reality, IEEE Computer Society, 2002, p. 97.
[27] P. Zikas, V. Bachlitzanakis, M. Papaefthymiou, G. Papagiannakis, A mobile, AR inside-out positional tracking algorithm, (MARIOPOT), suitable for modern, affordable cardboard-style VR HMDs, in: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection, Springer International Publishing, Cham, 2016, pp. 257–268.
[28] A. Vlachopoulos, Excavation at Vathy, Astypalaia ("Ανασκαφή στο Βαθύ Αστυπάλαιας")., Proceedings of the Archaeological Society at Athens, 2012, pp. 115–123.