Journal of Computational Science 32 (2019) 36–43

Contents lists available at ScienceDirect

Journal of Computational Science

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

Automated identification and deep classification of cut marks on bones and its paleoanthropological implications

Wonmin Byeon[a,f,1], Manuel Domínguez-Rodrigo[b,c,d,*,1], Georgios Arampatzis[a,e,1], Enrique Baquedano[b], José Yravedra[b,d], Miguel Angel Maté-González[g], Petros Koumoutsakos[a,e]

[a] Computational Science and Engineering Laboratory, ETH Zurich, CH-8092, Zurich, Switzerland
[b] Institute of Evolution in Africa (IDEA), University of Alcala de Henares, Covarrubias 36, 28010 Madrid, Spain
[c] Real Complutense College at Harvard, 26 Trowbridge Street, Cambridge, MA 02138, USA
[d] Department of Prehistory, Complutense University, 28040 Madrid, Spain
[e] Collegium Helveticum, 8092 Zurich, Switzerland
[f] Learning and Perception Research, NVIDIA, Santa Clara, CA 95051, USA
[g] Department of Cartography and Terrain Engineering, Polytechnic School of Avila, University of Salamanca, Hornos Caleros 50, 05003 Avila, Spain

ARTICLE INFO

Article history:
Received 15 September 2018
Received in revised form 29 January 2019
Accepted 22 February 2019
Available online 26 February 2019

Keywords:
Cut mark
Trampling
Deep learning
Machine learning
Paleoanthropology
Taphonomy

ABSTRACT

The identification of cut marks and other bone surface modifications (BSM) provides evidence for the emergence of meat-eating in human evolution. This most crucial part of taphonomic analysis of the archaeological human record has been controversial due to highly subjective interpretations of BSM. Here, we use a sample of 79 trampling and cut marks to compare the accuracy in mark identification on bones by human experts and computer trained algorithms. We demonstrate that deep convolutional neural networks (DCNN) and support vector machines (SVM) can recognize marks with accuracy that far exceeds that of human experts. Automated recognition and analysis of BSM using DCNN can achieve an accuracy of 91% of correct identification of cut and trampling marks versus a much lower accuracy rate (63%) obtained by trained human experts. This success underscores the capability of machine learning algorithms to help resolve controversies in taphonomic research and, more specifically, in the study of bone surface modifications. We envision that the proposed methods can help resolve on-going controversies on the earliest human meat-eating behaviors in Africa and other issues such as the earliest occupation of America.

1. Introduction

The emergence of meat-eating is one of the most highly debated issues in human evolution. It has been linked to encephalization, stone tool use and other major changes in the structural behavior of early hominins (see review in [1]). Recent discoveries argue that both stone tools and meat-eating could have potentially occurred more than one million years before the earliest evidence of encephalization [2,3]. If true, this will constitute a paradigm shift on our current interpretations on how these behaviors emerged during human evolution. This emphasizes the importance of correctly identifying the bone surface modifications that support such an early meat-eating behavior. The correct interpretation of these modifications, and more specifically of cut marks, is also of great relevance for understanding other important early hominin behavioral features, including early archaeological site function, meat used as a regular or a fallback food, and hunting or scavenging as methods of carcass acquisition. Inference of other social components could be achieved through the analysis of the butchering process (e.g., carcass disarticulation patterns). In fact, the use of cut marks in the fossil record has enabled the identification of specific butchering behaviors by early Pleistocene hominins and their carcass acquisition strategies in the earliest stages of evolution of *Homo* (see review in [4]).

The uncertainty surrounding the spatial association of stone tools and faunal remains during palimpsest formation, dictates that only anthropogenic bone surface modifications can be used effectively to link functionally lithics and bones. Hence, there is great need to accurately identify cut marks and to distinguish them from modifications created by other non-hominin agents, such as trampling or sediment abrasion. Claims have been made about

* Corresponding author at: Institute of Evolution in Africa (IDEA), University of Alcalá de Henares, Covarrubias 36, 28010 Madrid.
E-mail address: m.dominguez.rodrigo@gmail.com (M. Domínguez-Rodrigo).
[1] These authors contributed equally.

https://doi.org/10.1016/j.jocs.2019.02.005
1877-7503/

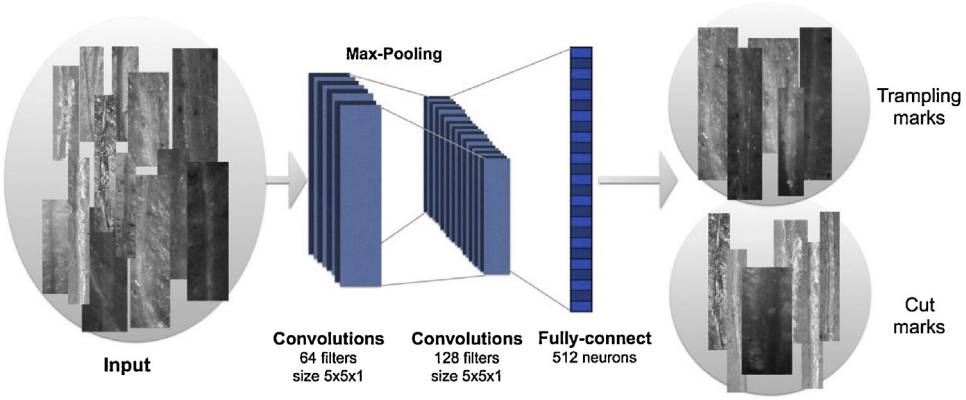


Fig. 1. Convolutional Neural Network Architecture: The network includes two convolutional with Rectified Linear Unit (ReLU) and Max-Pooling (MP) layers. The output of the last MP layer is connected to one fully-connected layer, then the network outputs the final classification.

the Pliocene antiquity of carcass processing using purported cut-marked bones in Dikika (Ethiopia) [2] or in the Plio-Pleistocene boundary in Quranwala (India) [5]. The claims made based on these discoveries remain controversial [6], as the purported cut-marked fossils from both sites also bear clear evidence of trampling or sedimentary abrasion [7].

The importance of a correct identification of bone surface modifications is not restricted to the earliest periods in human evolution. Recent claims have also been made for a 30,000-year-old presence of humans in the American continent based on the presence of cut-marked bones in old fossiliferous deposits (e.g., [8–11]). This contradicts current evidence of the earliest presence of humans in the continent at a much later age. Similar to the findings in Dikika and Quranwala, the Arroyo del Vizcaino site (Uruguay) also contains abundant conspicuous evidence of trampling and/or sedimentary abrasion [9–11].

Taphonomists broadly disagree on how cut marks could be properly identified. A series of microscopic criteria were experimentally defined [12–14], but there is substantial disagreement on how these criteria should be interpreted by individual researchers [13,15]. Recently, it has been documented that the interpretation of some of these variables remains subjective, hampering an objective scientific approach to cut mark identification and replication [44]. The lack of objective methods underscores the fact that the present identification of BSM (namely, cut marks) depends strongly on the subjective assessment and knowledge of each researcher [44]. Cut mark identification is, thus, commonly carried out outside replicable hypothesis-testing frameworks.

In order to introduce objectivity in cut mark identification, we introduce automatic image classification by machine learning algorithms. Algorithms such as Convolutional Neural Networks (CNN) [16–21] and Support Vector Machine (SVM) [22–24], have been shown to match and even exceed human performance in image and pattern recognition. Their use for image processing and classification is greatly facilitated by access to open source software packages such as Neuroph of TensorFlow [25–27].

In the present study the sample size (79 marks) was kept intentionally short in order to maximize human expert scores (the larger the sample the higher the identification failure rates by humans) and to minimize the computer's accuracy (larger sample sizes lead to better training and higher identification rates) [22–27]. Despite this initial advantage for human experts, we find that computer algorithms are more successful in the task of mark identification. The proposed CNN and SVM methods enable taphonomists to perform BSM identification in a more objective way than has ever been possible.

We present results that demonstrate the superiority of machine learning algorithms in identifying BSM over "subjective" assessment by several human experts. The present study, using modern bones, suggests that cut marks, like other taphonomic entities, are subject to morphological evolution, through a palimpsestic multiple-agent processes and an interplay between stasis and change [28–30]. This approach may enable the objective resolution of many cut mark-related controversies, including whether hominin butchering behaviors are identifiable in Pliocene fossils and whether the Americas were occupied by humans more than 30,000 years ago, given the presence of bones bearing BSM which could be interpreted as purported cut marks.

2. Methods

2.1. Methodological description of the sample

A selection of 79 experimental marks was used as the training set (SI). These were composed of 42 trampling marks and 37 cut marks. Trampling marks were created by using four types of sediments: fine-grained (0.06–0.2 mm), medium-grained (0.2–0.6 mm) and coarse-grained (0.6–2.0 mm) sand, as well as a combination of the previous sand types over a clay substratum, and granular gravel (>2.0 mm). These marks were selected from the trampling experiment reported by Domínguez-Rodrigo et al. [13] and they include all the variety of abrasive sediment particles (other than large pebble gravel grains ranging between 4–6 mm) potentially creating trampling marks in natural settings.

The set of cut marks was made using quartzite flakes as reported in Domínguez-Rodrigo et al.'s [13] experimental sample. The 37 cut marks were made with simple flakes ($n = 10$) and retouched flakes ($n = 27$). Although these marks are somewhat dissimilar they share far more similarities when compared to other non-anthropogenic marks. Retouched flakes were chosen as their resulting marks are statistically non-differentiable from cut marks created by natural rock flakes, since both are caused by irregular edges as opposed to the straighter edge of simple flakes [31]. One of the most polemic interpretations of the earliest purported cut marks (e.g., in Dikika) is that they may have been made with natural rocks. We employ marks made with retouched flakes to model the assumption that cut marks were created with modified tools for later periods. Both sets of cut marks were lumped together to discriminate between trampling and cut marks.

The selected marks were photographed in a standardized way under the binocular microscope at 30x. Images were then converted to grayscale. Further transformation of the original files was carried out as explained in the computing methods below.

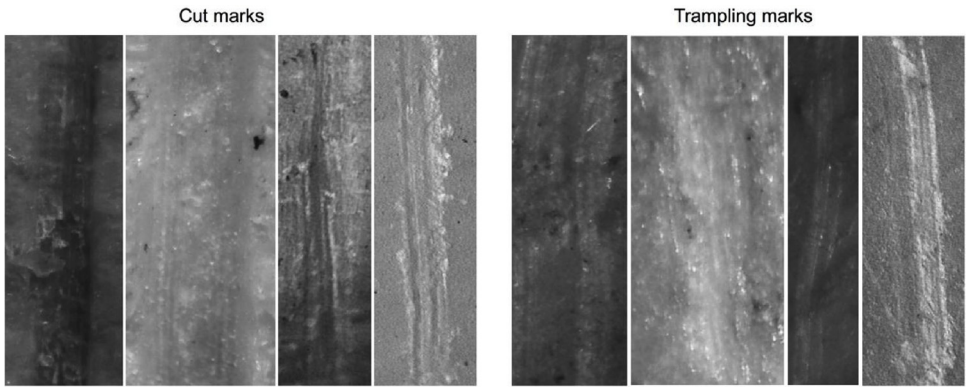


Fig. 2. Examples of the challenging images depicting trampling and cut marks, which show different contrast and lighting. Despite this, CNN identified >90% of marks.

2.2. Image pattern recognition methods

2.2.1. Deep learning-based method

Convolutional Neural Networks (CNNs) are potent deep learning-based methods for image classification [19,32]. The networks receive an image as input and transforms it via several hidden layers. Each hidden layer includes convolutional, Rectified Linear Unit (ReLU), Max-Pooling (MP), and Fully-Connected (FC) layers. The convolutional layer computes the weights of a grid of neurons which are connected to the local regions (kernels) of each pixel. The ReLU layer contains the nonlinear activation function:

$$f(x) = \max(0, x).$$

The MP layer takes the maximum value from the rectangular region which down-samples and compresses the information along the spatial dimensions. Finally, the FC layer has fully-connected neurons from the activations of the previous layer. These connections are weight matrix multiplications with biases followed by the ReLU activation function. The final layer is the softmax function which computes the normalized posteriors. The networks is trained to minimize the cross-entropy loss between the output and the true distribution. The architecture of Convolutional Neural Networks is summarized in Fig. 1.

2.2.2. Support Vector Machine (SVM) method

In this approach we classify the images by training a binary SVM classifier [33]. First a vocabulary of visual words is created from images in the training set using the Bag of Words (BoW) technique [34]. In this method features are extracted from images and the feature space is clustered. The vocabulary is composed by the different clusters. The images in the data set are encoded using the BoW vocabulary and an SVM is trained on the encoded set. In case the data are not separable, an appropriate kernel can be introduced transforming the data and making them separable in the new space.

2.2.3. Testing by human experts

Three experienced taphonomist with 7–20 years of practical experience on bone surface modifications using modern (i.e., controlled) and fossil assemblages were selected to independently identify mark types in the image set described above. Their results were compared first among themselves and then with those provided by the computer vision approach. The three experts were trained by one of the senior authors (MDR) of this paper.

2.3. Experiments

2.3.1. Convolutional neural networks (CNNs)

2.3.1.1. Pre-processing. The size of original images in the dataset were heterogeneous; the width ranged between 573 and 1350 pixels and the height ranged between 71 to 375 pixels. To speed up the training and provide a fixed size input to the network, the image resolution was resized to 180x520 pixels. The image values were normalized between 0 and 1 by the min-max scaling:

$$X'_i = \frac{X_i - \min(X)}{\max(X) - \min(X)},$$

where X_i is the original value and X'_i is the normalized value at the position i. No other pre-processing was performed, although the illumination and the quality of the images in the dataset are very diverse (see Fig. 2). To normalize the brightness and enhance the contrast of the images, the histogram equalization algorithm from the OpenCV library is applied. The method rearranges the original distribution of the pixel values to a wide and a flat distribution. The purpose of this process is to stretch out the pixel value range, to enhance the contract or illumination when the image is too dark or bright.

2.3.1.2. Architecture. The input layer receives an image with the size 520×180. To find the optimal architecture, two to four convolutional layers with MP layers and one to two FC layers with various sizes of neurons, and filter size between three to seven were tested. Exponential Linear Units (ELU) activation [35] was also tested in addition to ReLU, but it yielded no performance difference.

Finally, two convolutional layers and one FC layer were used as a final model. The first and second convolutional layers have 64 and 128 convolutional kernels. The process was followed by ReLU activation and MP. The size of the kernels for convolutional layers was $5 \times 5 \times 1$ with a stride 1 and for max-pooling layers it was $2 \times 2 \times 1$ with a stride 2. In the last layer, one fully-connected layer with 512 neurons were connected to the output layer consisting of a binary classification.

2.3.1.3. Training. The network was trained by Adaptive Moment Estimation (Adam) [36]. The learning rate was initialized to 0.001 with an exponential decay with a rate of 0.99. L2-weight decay and dropout [37] were used as regularizers in the FC layer. The constant for the L2-weight decay was $5e{-}4$, including 20 of dropout. The weights were randomly initialized with zero mean and a 0.1 standard deviation, and the biases were initialized to the constant value 0.1. The Tensorflow GPU framework [34,36] was used for the CNN experiments.

2.3.2. Support vector machine (SVM)

2.3.2.1. Pre-processing. The images were normalized such that they all have zero mean and standard deviation one.

2.3.2.2. Architecture. The implementation of the SVM approach was done in Matlab using the Computer Vision System Toolbox.

The key-points in the images were selected from points either on a regular grid or detected with a SURF detector. In both cases, the vocabulary was created from all the images in the training set using features extracted with the SURF extractor [38,39]. A grid step size of 12×12 and a vocabulary size of 600 was yielding results comparable to those of the SURF detector with a vocabulary size of 400. For the SVM we choose polynomial and Gaussian kernels and the hyper parameters of the kernels (order and box constraint for the polynomial, scaling and box constraint for the Gaussian) are being optimized. The box constraint parameter is related to the penalization of misclassifications in the case of non-separable data sets. We refer to the help page of the Matlab function 'fitcsvm' for more details on the hyper-parameters and the optimization procedure.

2.3.2.3. Training. The hyperparameters were optimized using Bayesian optimization [39] and the Covariance Matrix Adaption Evolution Strategy (CMA-ES) algorithm [40] produced similar results.

2.3.3. Dataset

The experimental marks dataset consists of 79 gray scale images and is divided into a training and a test set [13]. The test set consists of 10 images from each category (cut and trampling) and the rest of images were used for training. The accuracy of the classifier is estimated by averaging the accuracy over 60 independent trainings and the mean, as well as, the standard deviation of the accuracy is reported. Finally, the classifier is trained using the whole experimental dataset.

2.3.4. Evaluation

For both, CNN and SVM classifiers, the accuracy of BSM identification was averaged over 60 models trained by randomly partitioning the experimental dataset. This evaluation technique gives reliable validation of the model; the performance was not biased by a particular partition of training and testing sets. In our experiments, SVM was less stable than CNNs depending on the partition of the samples.

In order to compare the classifiers, three commonly used measures were computed, accuracy, specificity, and sensitivity. For the computation of the measures in binary classification, one category is chosen and marked as the 'positive' class and the other as the negative. Then we define the following sets,

- true positive (tp): elements belonging in class 'positive' and classified as positive,
- false positive (fp): elements belonging in class 'negative' and classified as positive,
- true negative (tn): elements belonging in class 'negative' and classified as negative,
- false negative (fn): elements belonging in class 'positive' and classified as negative.

Finally, the performance measures are defined as,

$$accuracy = \frac{|tp|}{|tp| + |fp|},$$

$$sensitivity = \frac{|tp|}{|tp| + |fn|},$$

$$specificity = \frac{|tp|}{|tp| + |fp|},$$

where $|A|$ denotes the number of elements belonging in set A. In our evaluation, cut marks were set to 'positive' marks and trampling marks to negative marks.

Table 1
Sensitivity and specificity of the CNN and SVM classifiers.

	Sensitivity	Specificity
CNN	90% (sd = 7.8)	90% (sd = 9.6)
SVM (GRID)	80% (sd = 13)	80% (sd = 13)
SVM (SURF)	84.5% (sd = 12.2)	82% (sd = 11.5)

Table 2
Classification probabilities for experts, CNN and SVM using SURF points for the images presented in Fig. 3. The cells with red color have been misclassified.

image number	experts	CNN	SURF
1	2/1	100%	54.3%
2	2/1	93.5%	95.7%
3	0/3	100%	76.5%
4	3/0	100%	96.7%
5	2/1	100%	96.6
6	2/1	98.6%	75%
7	2/1	76.5%	76.8%
8	2/1	89.9	81.6%

3. Results

For evaluation, 20 experimental marks are tested to compare the performance of CNNs, SVMs, and human experts.

The best CNNs architecture identified marks correctly with a 91% mean accuracy (sd = 5.3). Note that cut marks and trampling marks have balanced classification accuracy unlike SVM and human experts: 90% mean accuracy (sd = 10.6) for cut marks and 90% mean accuracy (sd = 8.0) for trampling marks.

The best accuracy obtained by SVMs is 81.5% (sd = 7.5) using the grid method to select points and 83% (sd = 8) using the SURF selector. For the SURF detector, the mean accuracy of trampling mark identification is 80% (sd = 80) and the mean accuracy of cut mark identification is 84.3% (sd = 13).

The results for the sensitivity of the CNN and SVM classifiers are being summarized in Table 1.

In contrast to the high accuracy of mark identification by CNN and SVM, the human experts identified a substantially lower number of marks. The three taphonomists produced similar identification rates; that is, their correct and incorrect identifications were similar among the three. A chi-square test showed that their identifications are statistically indistinguishable ($X^2 = 1.7577$; $p = 0.780$). They identified correctly a higher percentage of cut marks (mean = 66%; range = 64–70%) than of trampling marks (mean = 60%; range = 52–65%). Their overall correct identification rate averaged 63% of marks. We note that the nature of trampling marks implies larger variation in particle size and directions of the marks.

We visualize the performance of the classifier by presenting in Fig. 3 various cases of images (see also Table 2). For each image, we show the original image (left) the CNN gradients (middle) and the SURF points (right).

For the CNN, the gradient visualization is performed using the Gradient-weighted Class Activation Mapping (Grad-CAM) [41]. The Grad-CAM is an algorithm that creates a heatmap using the gradients of the predicted class with respect to the final convolutional feature map. Here, we show the heatmap overlapped with the

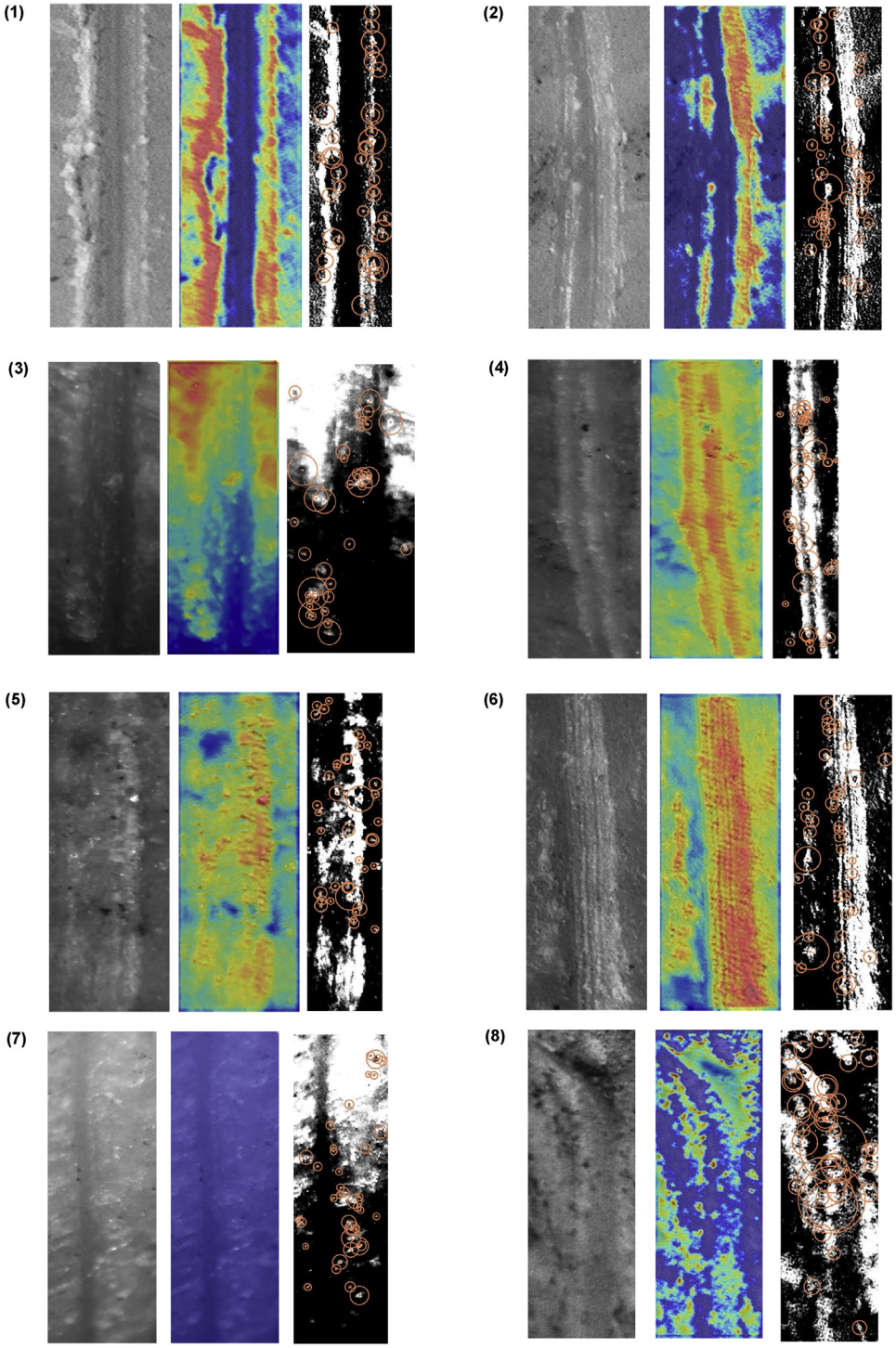


Fig. 3. Cut marks (1–3, 7) and trampling marks (4–6, 8). Images (7) and (8) were misclassified by CNN and image (6) by SVM. For each figure: original (left), CNN Grad-Cam [41] (middle) and SURF points (left).

original (input) images. This visualization technique highlights the important regions for the prediction; red regions are more important and blue less. The Grad-CAM results show that CNN learns particular patterns for cut marks and trampling marks to make a correct identification.

For the SVM, we show the SURF points [38,39] as dots. The circle around the point shows the importance of the SURF point. Here, we show the 40 most important points.

There was no correlation between years of experience and accuracy among the taphonomy experts. As a matter of fact, the most senior expert produced the lowest accuracy. For all the images of Fig. 3 we have the comments from the expert taphonomists:

(1) The first cut mark was identified correctly by two experts because it shows a straight continuous trajectory of microstriations, the apparent V-shaped section of the groove and the conspicuous asymmetrical flaking of the shoulder. (2) The second cut mark was also identified correctly by two experts because the section is also V-shaped and the microstriations are continuous although the trajectory is slightly curved. The shoulder effect is asymmetrical and very conspicuous and flaking is taking place along the

mark shoulder. (3) The last cut mark was misidentified by all the experts, probably because the groove is very broad and the microstriations are not marked and continuous. (4) The first trampling mark, was correctly identified by all experts because the trajectory is wavy and there is a lack of noticeable flaking on the mark shoulder. Microstriations are also wavy and there is an absence of flaking along the groove trajectory. (5) The second trampling mark, was also correctly identified by two experts for the same reasons as in the mark (4). In this case, the broad groove section and the sinuosity of the mark trajectory and its microstriations are typical of trampling marks. (6) The last trampling mark, was correctly identified by two experts because of its broad section and the lack of adjacent flaking on the shoulder edges. (7) Two experts correctly identified this as a cut mark because of its narrow and deep section with continuous microstriations. (8) This was also correctly identified by two experts because of its broad section, very little marked microstriations and because of the lack of flaking on the edges or the presence of any other shoulder effect. Its markedly curved trajectory is also typical of trampling marks.

It is important to emphasize that the mark (3) is misclassified by all human experts but correctly classified by CNN and SVM. Also, all correctly-classified examples by CNN has high confidence compared to the misclassified ones. This is not the case for SVM, that doesn't show the same stability.

For the CNN misclassified examples of Fig. 3, Grad-CAM results shows that CNN found unclear patterns due to its blurry input images, but human experts identified them correctly. Overall, we found that CNN learns particular patterns to identify the marks as human experts do. As CNN identifies > 90% of marks while the human experts identify 61%, these patterns found by CNN play a major role to identify the marks. This verifies the importance of this study.

4. Discussion

A few years ago, a multivariate analysis of trampling and cut marks using a set of microscopic variables was argued to yield an accuracy of >80% in the identification of marks [13]. However, this high-accuracy classification rate was automatically obtained by a discriminant analysis. It did not test the performance of the human analyst. The study assumed that humans could objectively interpret the same categories in each variable and the discriminant analysis would automatically classify each mark. However, even in something as objective as defining a category within a variable there is wide variability of interpretations [44,13,15]. This creates an artificial divergence in the interpretations of bone surface modifications and their correct identification remains highly subjective.

The present study shows that expert taphonomists may be only moderately successful in identifying BSM in low (300×) magnification (accuracy = 63%). This is an insufficient success rate for elaborating reliable interpretations pending on correct identification of marks, such as those on the Dikika fossils or the Quranwala purported cut-marked bones. Ad hoc experimentation in the latter case may be as spurious as for the Dikika case [5]. Similarities between particular types of experimental and fossil marks, both in the case of the Quranwala and Dikika, are overshadowed by the plethora of other associated marks on those fossils that remain ignored and by other available experimental marks that are morphologically similar and objectively undifferentiated from those reported on those fossil bones [6].

The present findings indicate that machine learning algorithms can augment human capacity in the interpretation of bone surface modifications. However, the "objectivity" of the algorithms depends on the data training set. This quality depends not only on the within-sample diversity of each type of mark but also on the variety of marks types that are part of the training sample. All trampling marks used in this study are made with fine-grained, medium-grained and coarse-grained sand as well as small gravel [13]. Future analyses should include trampling marks created with pebble gravel, such as those reported by Domínguez-Rodrigo et al. [6]. Most cut marks used were made with flint and quartzite tools. Given the diversity in cut mark morphologies created by same tools made on different raw materials [42], marks created with other raw material types and using different tool types should also provide a wider panorama of the morphological diversity that each of these agents introduce.

We note that a larger image dataset may help refine the results presented herein. For instance, one of the most successful works for image classification using CNNs [19] used 1.3 million images for training, while the model in Ciresan et al. [18] for traffic sign recognition is trained on 25,350 images. The samples used in this work were very difficult to identify. The data samples have high variation in illumination, contrast and resolution and hard to visually identify (see Fig. 2). Minimal pre-processing (only image resizing and normalization) and no data augmentation were performed in model training. A wide variety of image processing and transformation methods are frequently used in the literature to increase the robustness of the model, especially for deep learning approaches [16–21]. Common image pre-processing such as histogram equalization and contrast normalization enhance the quality of images. Data augmentation includes random transformation, rotation, multi-scaling, and flipping, which diversify the dataset. Such techniques usually help to increase the classification performance. Despite not having applied any of these improvement sample methods, our results outperform human experts by a wide margin (almost 50%). This fact, regardless of the sample size used, underscores the drastic improvement in mark identification by machine learning methods compared to human experts.

This work showcases the capabilities of deep learning algorithm to resolve the highly-controversial issue of BSM identification in taphonomy. We have used different types of sand grains so that the trampling experimental sample used has reproduced the most common type of trampling marks usually encountered in the recent and fossil records. Likewise, by using simple and retouched quartzite flakes, this experimental set has also reproduced the most common form of butchering tools used in several early and middle Pleistocene sites (which is the main target of this study). Given that quartzite produces a more similar mark morphology to trampling marks than flint, the use of cut marks made with flint or highly-crystalline raw materials, such as obsidian, would only discriminate the resulting cut marks better from trampling marks [43].

The present approach provides an objective and accurate method to identify marks, and it can assist taphonomists further by accessing larger samples than those available in this study. In the process, researchers should be aware that the quality of their classifications will depend tightly on the patterns the algorithm has accessed via training. In this regard, the computer is like a human. If the algorithm is trained with an insufficient number of mark types and these do not represent the variability range of the population, the accuracy in the identification will be highly compromised.

5. Conclusions

We have implemented convolutional deep neural networks to identify cut and trampling marks on bones. The algorithms exhibit an accuracy that is almost 50% better than those produced by experienced taphonomists trained on BSM. The data and analysis presented here are introductory to the potential of machine learning algorithms in taphonomic research. The present methods are readily extensible to hundreds or thousands of images suggesting that taphonomic research can be dramatically improved by creating image databases of marks that may be curated by multiple agen-

cies. This would enable studies with an empirically well-founded graphic referential framework for the interpretation of past bone surface modifications. Once this referential database is available, taphonomists would be in a better position to interpret if controversial marks such as those from the Dikika, Quranwala or Arroyo del Vizcaino fossils are indeed anthropogenic. The present work demonstrates that machine learning may assist human knowledge in the taphonomic discipline and we envision that it will be broadly adopted by the community.

Method online

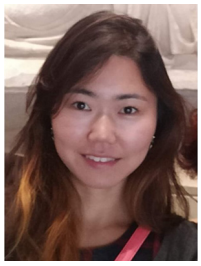
A step-by-step description and the code used can be found at https://github.com/cselab/cut-marks-classification.

Acknowledgements

This collaborative work was carried out with support from a Research Salvador Madariaga grant to MDR (Ministry of Education, Culture and Sport, Spain. Ref PRX16/00010). WB, GA and PK acknowledge support by the ERC Advanced Investigator award No 341117 (FMCoBe). MDR thanks D. Lieberman and the Human Evolutionary Biology Department at Harvard and the Royal Complutense College at Harvard, where this research was conducted. We also thank Lucía Cobo and Julia Aramendi for their help during the design of this pilot study and three taphonomic experts, who wish to remain anonymous in the present paper. We are indebted to the comments made by J. Heaton on a earlier version of this paper.

References

[1] M. Domínguez-Rodrigo, T.R. Pickering, The meat of the matter: an evolutionary perspective on human carnivory, Azania: Archaeol. Res. Afr. 3 (2017) 1–29.
[2] S.P. McPherron, Z. Alemseged, C.W. Marean, J.G. Wynn, D. Reed, D. Geraads, et al., Evidence for stone-tool-assisted consumption of animal tissues before 3.39 million years ago at Dikika, Ethiopia, Nature 466 (2010) 857–860.
[3] S. Harmand, J.E. Lewis, C.S. Feibel, C.J. Lepre, S. Prat, A. Lenoble, et al., 3.3-million-year-old stone tools from Lomekwi 3, West Turkana, Kenya, Nature 521 (2015) 310–315.
[4] M. Domínguez-Rodrigo, R. Barba, C.P. Egeland, Deconstructing Olduvai: A Taphonomic Study of the Bed I Sites, Springer Science & Business Media, 2007.
[5] A. Dambricourt Malassé, A.-M. Moigne, M. Singh, T. Calligaro, B. Karir, C. Gaillard, et al., Intentional cut marks on bovid from the Quranwala zone, 2.6 Ma, Siwalik Frontal Range, northwestern India, C. R. Palevol 15 (2016) 317–339.
[6] M. Domínguez-Rodrigo, T.R. Pickering, H.T. Bunn, Configurational approach to identifying the earliest hominin butchers, Proc. Natl. Acad. Sci. USA 107 (2010) 20929–20934.
[7] M. Domínguez-Rodrigo, L. Alcalá, 3.3-Million-year-old stone tools and butchery traces? More evidence needed, PaleoAnthropology 16 (2016) 46–53.
[8] R.A. Fariña, S.F. Vizcaíno, G. De Iuliis, Megafauna. Giant Beasts of Pleistocene South America, Indiana University Press, Bloomington, 2013, 416 pp.
[9] R.A. Fariña, P.S. Tambusso, L. Varela, M. Di Giacomo, M. Musso, A. Gascue, R. Bracco, Among others, cut-marks are archaeological evidence: Reply to "Archaeological evidences are still missing: Comment on Fariña et al. Arroyo del Vizcaíno Site, Uruguay" by Suárez et al, Proc. R. Soc. B 281 (1795) (2014) 20141637.
[10] R.A. Fariña, P.S. Tambusso, L. Varela, A. Czerwonogora, M. Di Giacomo, M. Musso, R. Bracco-Boksar, A. Gascue, Arroyo del Vizcaíno, Uruguay: a fossil-rich 30-ka-old megafaunal locality with cut-marked bones, Proc. R. Soc. B 281 (1774) (2014) 20132211.
[11] R.A. Fariña, Bone surface modifications, reasonable certainty and human antiquity in the Americas: the case of the arroyo del Vizcaíno site, Am. Antiq. 80 (1) (2015) 193–200.
[12] J.W. Fisher, Bone surface modifications in zooarchaeology, J Archaeol Method Theory 2 (1995) 7–68.
[13] M. Domínguez-Rodrigo, S. de Juana, A.B. Galán, M. Rodríguez, A new protocol to differentiate trampling marks from butchery cut marks, J. Archaeol. Sci. 36 (2009) 2643–2654.
[14] S. de Juana, A.B. Galán, M. Domínguez-Rodrigo, Taphonomic identification of cut marks made with lithic handaxes: an experimental study, J. Archaeol. Sci. 37 (2010) 1841–1850.
[15] G.F. Monnier, E. Bischoff, Size matters. An evaluation of descriptive criteria for identifying cut marks made by unmodified rocks during butchery, J. Archaeol. Sci. 50 (2014) 305–317.
[16] B. Fasel, Facial expression analysis using shape and motion information extracted by convolutional neural networks, in: Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing, IEEE, 2002, pp. 607–616.
[17] F.H.C. Tivive, A. Bouzerdoum, A face detection system using shunting inhibitory convolutional neural networks, in: IEEE International Joint Conference on Neural Networks (IEEE Cat No04CH37541), IEEE, 2004, pp. 2571–2575.
[18] D. Ciresan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 3642–3649.
[19] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25, Curran Associates, Inc., 2012, pp. 1097–1105.
[20] S. Hijazi, R. Kumar, C. Rowen, Using Convolutional Neural Networks for Image Recognition [Internet], 2015, Available at: http://www.multimediadocs.com/assets/cadence_emea/documents/using_convolutional_neural_networks_for_image_recognition.pdf.
[21] H. Yalcin, S. Razavi, Plant classification using convolutional neural networks, in: Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics), IEEE, 2016, pp. 1–5.
[22] C.J.C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (1998) 121–167.
[23] P.J. Phillips, Support vector machines applied to face recognition, in: M.J. Kearns, S.A. Solla, D.A. Cohn (Eds.), Advances in Neural Information Processing Systems 11, MIT Press, 1999, pp. 803–809.
[24] S.-W. Lee, A. Verri, Pattern Recognition with Support Vector Machines: First International Workshop, SVM. Niagara Falls, Canada, August 10, 2002. Proceedings 2003, Springer, 2002.
[25] M. Egmont-Petersen, D. de Ridder, H. Handels, Image processing with neural networks—a review, Pattern Recognit. 35 (2002) 2279–2301.
[26] J.S. Ueli Meier, Multi-Column Deep Neural Networks for Image Classification, 2012, Available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.367.484.
[27] M. Nielsen, Using Neural Nets to Recognize Handwritten Digits, Determination Press, 2015.
[28] S.R. Fernández López, Tafonomía y fosilización. eprints.ucm.es, 1999, Available at: http://eprints.ucm.es/21802/1/078_99_Tafonomia_y_Fosilizacion.pdf.
[29] S.R. Fernández López, Temas de tafonomía. eprints.ucm.es, 2000, Available at: http://eprints.ucm.es/22003/1/087_00_Temas_Tafonomia.pdf.
[30] S.R. Fernández López, Taphonomic alteration and evolutionary taphonomy, J. Taphon. 4 (2006) 111–142.
[31] M. Domínguez-Rodrigo, T.R. Pickering, H.T. Bunn, Experimental study of cut marks made with rocks unmodified by human flaking and its bearing on claims of 3.4-million-year-old butchery evidence from Dikika, Ethiopia, J. Archaeol. Sci. 39 (2012) 205–214.
[32] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE (1998) 2278–2324.
[33] I. Steinwart, A. Christmann, Support Vector Machines, Springer Science & Business Media, 2008.
[34] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, Workshop on Statistical Learning in Computer Vision, ECCV, Prague, 2004, pp. 1–2.
[35] D.-A. Clevert, T. Unterthiner, S. Hochreiter, Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs) [Internet], 2015, arXiv [cs.LG]. Available at: http://arxiv.org/abs/1511.07289.
[36] L. Rampasek, A. Goldenberg, TensorFlow: biology's gateway to deep learning? Cell Syst. 2 (2016) 12–14.
[37] N. Srivastava, G.R. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Over tting, 2014, Available at: http://jmlr.org/papers/volume15/srivastava14a/srivastava14a.pdf.
[38] M. Pelikan, D.E. Goldberg, E. Cantú-Paz, BOA: The Bayesian optimization algorithm of the 1st Annual Conference on . . .. dl.acm.org, 1999, Available at: http://dl.acm.org/citation.cfm?id=2933973.
[39] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. 110 (2008) 346–359.
[40] N. Hansen, S.D. Müller, P. Koumoutsakos, Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), Evol. Comput. (2003).
[41] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017.
[42] J. Yravedra, Maté-González MÁ, J.F. Palomeque-González, J. Aramendi, V. Estaca-Gómez, M. San Juan Blazquez, et al., A new approach to raw material use in the exploitation of animal carcasses at BK (Upper Bed II, Olduvai Gorge, Tanzania): a micro-photogrammetric and geometric morphometric analysis of fossil cut marks. Boreas, 2017.
[43] M.Á. Maté-González, J.F. Palomeque-González, J. Yravedra, D. González-Aguilera, M. Domínguez-Rodrigo, Micro-photogrammetric and morphometric differentiation of cut marks on bones using metal knives, quartzite, and flint flakes, Archaeol. Anthropol. Sci. (2016) 1–12.
[44] M. Domínguez-Rodrigo, P. Saladié, I. Cáceres, R. Huguet, J. Yravedra, A. Rodríguez-Hidalgo, P. Martín, A. Pineda, J. Marín, C. Gené, J. Aramendi, L. Cobo-Sánchez, Use and abuse of cut mark analyses: the Rorschach effect, J. Archaeol Sci. 86 (2017) 14–23.



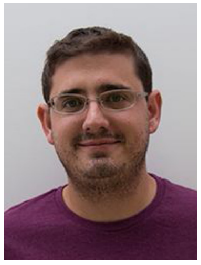
Wonmin Byeon is a researcher at NVIDIA Research in Santa Clara, US. Before joining NVIDIA, in 2017, she was a post-doctoral researcher at ETH Zurich and IDSIA, Switzerland, working with Juergen Schmidhuber and Petros Koumoutsakos. She received her PhD in Computer Science from Technical University Kaiserslautern, Germany in 2016. Her research interests are in the fields of deep learning and computer vision, especially with Multi-dimensional Long Short-Term Memory recurrent neural network (MD-LSTM) and Convolutional Neural Network (CNN) for high-dimensional data understanding. Email: wonmin.byeon@gmail.com

Manuel Domínguez-Rodrigo is co-director of the IDEA (Institute of Evolution in Africa) and professor of the Department of Prehistory, Ancient History and Archeology of the Complutense University. He has been co-director of the paleoanthropological projects of Peninj (Lago Natron) (1995–2005), Eyasi (2002–2006) and, currently, of the Olduvai Gorge (2006–present). He has published 8 books and more than 200 impact articles. He has been guest professor and researcher at the Universities of Harvard, Rutgers and St. Louis and the Royal Complutense College in Harvard (USA). His specialties are taphonomy and paleoanthropology. He is pioneering the application of high computing tools, such as algorithms of "machine learning" and "deep learning" or "computer vision" to the world of paleoanthropology. He is currently co-director of TOPPP (www.olduvaiproject.org). Email: manueldr@ghis.ucm.es.

Georgios Arampatzis works since January 2015 as a postdoctoral researcher at the Computational Science & Engineering Laboratory at ETH Zurich and since 2018 at Collegium Helveticum. He received his Bachelor, Master and PhD degrees from the Mathematics and Applied Mathematics Department at the University of Crete in 2006, 2011 and 2014, respectively. During 2014–2015 he worked as a postdoctoral researcher at the Mathematics and Statistics Department at the University of Massachusetts, Amherst. His research interests include numerical computations, Monte Carlo methods, spatio-temporal kinetic Monte Carlo, sensitivity analysis for stochastic processes, uncertainty quantification algorithms with applications in molecular dynamics, fluid dynamics and pharmacodynamics and differential privacy.

Enrique Baquedano is co-director of the IDEA (Institute of Evolution in Africa) and director of the Regional Archaeological Museum. He is also co-director of TOPPP (www.olduvaiproject.org) and the excavations of the Neanderthal site of Pinilla del Valle (Madrid). His doctoral research work focused on taphonomic and historiographic topics. He is a visiting professor at the University of Alcalá de Henares. He has published several impact articles and is an expert in disseminating heritage through a large number of exhibitions. He is the architect of the permanent exhibitions of the Regional Archaeological Museum of Madrid and the museums of Olduvai and the National Museum of Tanzania in Dar es Salaam (Tanzania).

José Yravedra is a professor at the UCM and a member of TOPPP since 2008. He is also the director of several research projects, such as: "The evolution of human behavior a million years ago in East Africa, reviewing the evidence of beds III and IV of the Olduvai Gorge. The exploitation of megafaunas in the African Lower Paleolithic, new perspectives from BK (Olduvai Gorge)". He has also directed other projects in the Iberian Peninsula and has been involved in more than 40 research projects in African and Iberian sites. On the other hand, he has more than 300 publications distributed in books such as "Taphonomy applied to Zooarchaeology", monographs and scientific articles published in international journals such as Nature Communications, Quaternary Science Reviews, Boreas, Journal of Human Evolution, Journal of Archaeological Science, Quaterary International etc. His lines of research are mainly Paleolithic, Taphonomy and Zooarchaeology. He is currently working very actively in taphonomy of carnivores with specialization on canids and felines, and in the creation of new taphonomic documentation techniques through the development of microtaphonomy, being a pioneer in the application of photogrammetry and geometric morphometry.



Miguel Ángel Maté-González is a PhD in Geotechnologies applied to Construction, Energy and Industry (USAL-2017). He is currently a teaching and research staff at the University of Salamanca. He has solved difficulties related to the use of geomatics sensors, representation techniques, 3D display, sensor calibration and has generated a new methodology for the analysis of 3D cut marks on bones. That methodology has been used in different archaeological and palaeontological projects and numerous scientific papers and has been presented at various international congresses.

Petros Koumoutsakos holds the Chair for Computational Science at ETH Zurich and serves as Fellow of the Collegium Helveticum. Petros is elected Fellow of the American Society of Mechanical Engineers (ASME), the American Physical Society (APS), the Society of Industrial and Applied Mathematics (SIAM). He has held visiting fellow positions at Caltech, the University of Tokyo, MIT and the Radcliffe Institute of Advanced Study at Harvard University. He is recipient of the Advanced Investigator Award by the European Research Council and the ACM Gordon Bell prize in Supercomputing. He is elected Foreign Member to the US National Academy of Engineering (NAE). His research interests are on the fundamentals and applications of computing and data science to understand, predict and optimize fluid flows in engineering, nanotechnology, and medicine.