Revisiting giraffe photo-identification using deep learning and network analysis

Vincent Miele, Gaspard Dussert, Bruno Spataro, Simon Chamaillé-Jammes, Dominique Allainé, Christophe Bonenfant

Revisiting giraffe photo-identification using deep learning and network analysis

Vincent Miele[1], Gaspard Dussert[1], Bruno Spataro[1], Simon Chamaillé-Jammes[2,3,4], Dominique Allainé[1,4], Christophe Bonenfant[1,4]

[1] Université de Lyon, F-69000 Lyon; Université Lyon 1; CNRS, UMR5558, Laboratoire de Biométrie et Biologie Évolutive, F-69622 Villeurbanne, France. E-mail: vincent.miele@univ-lyon1.fr

[2] CEFE, Univiversité Montpellier, CNRS, EPHE, IRD, Université Paul Valéry, Montpellier, France

[3] Mammal Research Institute, Department of Zoology & Entomology, University of Pretoria, Pretoria, South Africa

[4] LTSER France, Zone Atelier "Hwange", Hwange National Park, Bag 62, Dete, Zimbabwe - CNRS HERD (Hwange Environmental Research Development) program

Abstract

An increasing number of research programs rely on photographic capture-recapture (vs. direct marking) of individuals to study distribution and demography within animal populations. Photo-identification of individuals living in the wild is sometimes feasible using idiosyncratic coat or skin patterns, like for giraffes. When performed manually, the task is tedious and becomes almost impossible as populations grow in size. Computer vision techniques are an appealing and unavoidable help to tackle this apparently simple task in the big-data era. In this context, we propose to revisit giraffe re-identification using convolutional neural networks (CNNs).We first developed an end-to-end pipeline to retrieve a comprehensive set of re-identified giraffes from about $4,000$ raw photographs. To do so, we combined CNN-based object detection, SIFT pattern matching, and image similarity networks. We then quantified the performance of deep metric learning to retrieve the identity of known and unknown individuals. The re-identification performance of CNNs reached a top 5 accuracy of about 90%. Fully based on open-source software packages, our work paves the way for further attempts to build CNN-based pipelines for re-identification of individual animals, in giraffes but also in other species.

Keywords: animal identification, SIFT, deep metric learning, image similarity networks, open-source

Introduction

In many respects, population and behavioural ecology have immensely benefited from individual-based, long term monitoring of animals in wild populations [1, 2]. At the heart of such monitoring is the ability to recognize individuals. This is often achieved by actively marking individuals, such as deploying ear-tags or leg rings, cutting fingers or feathers, or scratching scales in reptiles. In some species, however, individuals display natural marks that make them uniquely identifiable. Many large African mammals such as leopard (*Panthera pardus*), zebra (*Equus sp.*), kudu (*Tragelaphus strepsiceros*) or wildebeest (*Connochaetes taurinus*) all present idiosyncratic fur and coat patterns particularly useful for non-invasive and reliable recognition of individuals. This is the case for the iconic but endangered giraffe (*Giraffa camelopardalis*), for which mark-resight (MS) surveys would allow understanding the demographic processes underlying the dynamics of its populations, possibly helping to set conservation policies. Individual identification of giraffes has long been known to be feasible from comparisons of the distinctive coat patterns of individuals [3]. As the number of giraffes to identify increases, people-based visual comparisons of pictures can rapidly become overwhelming. With the move to digital technologies (namely digital cameras and camera traps), the problem becomes even more acute as the number of pictures to process can easily reach the thousands or ten of thousands.

Over the last decade, the use of computer vision rapidly spread into biological sciences to become almost unavoidable in animal ecology for many repetitive tasks [4]. In a seminal publication, Bolger and colleagues [5] first presented computer-aided giraffe photo-identification, soon followed by other similar studies on giraffe as well [6, 7]. The underlying computer technique is a feature matching algorithm, the Scale Invariant Feature Transform operator (SIFT; [8]), where each image is associated to the k-nearest best matches. The current use of SIFT for ecologists requires human intervention to validate the proposed candidate images within a graphical interface [9]. As SIFT features cover whole images, two images can be considered as similar not only because of similar giraffe coat patterns but also because of similarities in the backgrounds (similar trees for instance), which could lead to false positive matches. For the best results with computer vision, all images have to be cropped before, so that only the giraffe flank appears on the images, hence excluding most of the neck, head, legs and background. Until now, this cropping operation was most often done manually [6], despite being a highly time-consuming task when processing thousands of images. Recently, Buehler and colleagues [7] have however developed a procedure based on Histogram of Oriented Gradients (HOG) to automatise photograph cropping.

In the meantime, the Deep Learning (DL) revolution was underway in computer vision, showing breakthrough performance improvements [10]. In particular, convolutional neural networks (CNNs) are now the front-line technique to deal with the large range of image processing questions in ecology and environmental sciences [11]. In particular, many recent studies tackle the problem of re-identification using CNNs, which has been mostly developed and extensively used for humans [12]. Technically, re-identification consists in using a CNN to classify images of different individuals, some of them being not necessarily seen before, *i.e.* unknown individuals. However, despite the availability of proven and efficient techniques [13] and several successful attempts to apply the method to non-human species [14, 15, 16, 17, 18, 19, 20, 21, 22, 23], re-identification remains a challenging task when applied to animals in wild population where re-observations are limited *sensu largo* [24].

In practice, current CNN-based approaches have to be tailored to the needs of field ecologists interested in using them for individual recognition. For instance, batches of new images are regularly added to the reference database following yearly fieldwork sessions. Also, in many situations, new individuals may have been born, and the study population may be open, i.e. immigration can occur.

Therefore, we expect the re-sighting of known individuals, as well as the observation of individuals never seen before. In other words, this sampling design implies to solve the re-identification in a mixture of known and unknown individuals. Chen and colleagues [23] referred to this problem as the "open set" identification problem, and they proposed to identify images from unknown individuals and to assign them a single "unknown" label. Here, we moved one step further and evaluated the possibility to build a model capable of identifying all individuals, be they known or unknown.

A classical CNN classifier can re-identify already known individuals (usually with a *softmax* last layer) but will fail to identify new individuals. Indeed, the number of predicted classes must match the number of known individuals. Therefore, we crucially need a CNN-based approach that can retain the power of the features learnt in a CNN while, at the same time, allow for the identification of individuals unknown at the time of the analysis. We propose to rely on deep metric learning [DML 25] as an ideal candidate to solve the "open set" identification problem: DML consists in training a CNN model to embed the input data (input images) into a multidimensional Euclidean space such that data from a common class (for instance, images of a given individual) are much closer, in terms of Euclidean distance, than with the rest of the data. Retrieving the individuals (known as well as unknown) consists in relying on the Euclidean distance computed for any pair of images. Finally, a suitable machine learning algorithm will retrieve the individuals.

Here we addressed the problem of giraffe photo-identification with an updated, open-source, and end-to-end automatic pipeline. In a first step, we applied state-of-the art techniques for object detection with CNNs [26] to automatically crop giraffe flanks of about 4,000 raw photographs shot in the field. Indeed, the most recent CNN approaches clearly outperformed the HOG approach [27]. Second, following Bolger and colleagues [5], we used the SIFT operator to calculate a numeric distance between any pair of giraffe flanks. From the $n \times n$ calculated distances, we built an image similarity network [28] and applied network clustering to retrieve different clusters of images coming from different individual giraffes, hence removing any human intervention in the process. However, we manually validated a subset of our results to build a ground-truth data set of different individuals ($n = 178$) to train our CNN. Finally, we evaluated the predictive accuracy of our CNN-based metric learning approach with a 5-fold cross-validation procedure.

Material and Methods

Photograph database

We carried out this study in the northeast of Hwange National Park (HNP), Zimbabwe. HNP park covers a 14,650 km^2 area [29]. The giraffe sub-species currently present in HNP could be either *G. c. angolensis* or *G. c. giraffa* according to the IUCN [30]. Here we used data from a regular monitoring conducted between 2014 and 2018. Each year, daily for at least three consecutive weeks, we drove the road network available within ¡60km of the HNP Main Camp station, and took photographs of every giraffe encountered. Pictures were taken with 200mm to 300mm lenses mounted on Nikon DSRL cameras (sensor resolution ranged between 16 and 40 Mpx). Importantly, we filtered sequences of very similar photographs occurring with the camera burst mode in the same second, and retained one single photograph per sequence. Overall, we shot $n = 3{,}940$ photographs.

Image cropping with CNN and transfer learning

We relied on convolutional neural networks (CNNs) to detect one or several giraffes in each photograph. For an efficient detection and classification, a CNN has to be trained on a huge amount of images (usually > millions of images) to capture the most discriminant features associated with

each class. Because of our limited amount of photographs, we relied on *transfer learning* [31], a specific method aiming at training a CNN on a small number of images. In transfer learning, we do not start the CNN training "from scratch" with some random model parameters, but use the parameters of another model pre-trained for a task presenting some similarities with the task of interest. For giraffe detection, we used a pre-trained model for object detection that can deal with a few mammalian species, even if those species differed from the one of interest. This approach works because the pre-trained model has already learnt a wide range of relevant and generic features, that we can reuse as a starting point for our learning problem of finding giraffe flanks in photographs. Thanks to this prior information, transfer learning can deal with a small dimension data set (down to a few hundreds of images per class).

A range of cutting edges tools are now available to take advantage of CNN in the context of object detection for animal detection [32, 33, 34]. In particular, RetinaNet [26] is a CNN-based object detector able to detect a series of predefined object classes (*e.g.* different animal species) and that returns the coordinates of a bounding box around these objects as well as a confidence score. These two steps are performed at the same time with a single CNN, which makes RetinaNet a *one-stage* detector as opposed to two-stage detectors for which a first CNN search for regions containing a potential object and a second CNN classify these regions. Two-stage detectors are known to achieve better performance in practice but are slower than one-stage detectors [35]. However, RetinaNet proposes a new technique that better manage non informative objects' background with similar performance compared to two-stage detectors while being much faster [26]. Finally, it is known that the more heterogeneous the training data set is (various positions, backgrounds, scale or lighting), the most efficient a CNN is [36], so we used data augmentation (flipping, rotation and color changes) to enhance our model performance.

We manually prepared our training data set by cropping bounding boxes around giraffe flanks, excluding most of the neck, head, legs and background, with the `labelImg` open source program for image annotation (`https://github.com/tzutalin/labelImg`). We obtained 469 bounding boxes associated to 400 photographs. We performed transfer learning with RetinaNet to detect a single object class, the giraffe flank, from a pre-trained model shipped with RetinaNet, that is a ResNet50 backbone trained on the COCO dataset (80 different classes of common objects including giraffes among a few other animal species [37]). We trained the model with 30 epochs of 100 batches of size 2. Training took approximately 30 minutes on a Titan X card. Our pipeline was based on the Keras implementation of RetinaNet available at `https://github.com/fizyr/keras-retinanet`.

Distance estimation between giraffe flank images

Our re-identification pipeline is based on the computation of distance between pairs of images displaying giraffe flanks. To build a reference data set and to gain some insights on the potential added-value of deep-learning approaches, we used either distances based on the SIFT algorithm, currently the most commonly used traditional computer vision approach in our context, or distances based on a CNN approach. These distances were then used to build a network of photographs and identify clusters of photographs of the same individual giraffe.

Using the Scale Invariant Feature Transform operator

We built on Bolger and colleagues [5] to achieve pattern matching between giraffe flanks with the Scale Invariant Feature Transform operator (SIFT; [8]). The SIFT algorithm extracts characteristic features in photographs called *key points* that are invariant with respect to scale and orientation. Comparing two photographs, pairs of matching key points (i.e. having similar characteristics) are

retrieved and ranked by distance (Euclidean distance between their feature vectors). Here, we selected the 25 closest pairs of key points. However, for better results, we had to assess the extent to which matching key points were coherent in the two giraffe flanks, *i.e.* if their location matched on the giraffe body. To find out relevant cases where matching key points were actual matches of coat patterns, we superimposed key points extracted from a pair of giraffe photographs with a geometrical transformation called *homography*. An homography is a perspective transformation between the two planes, with here one plane per image.The homography consists in finding the optimal transformation such that the key points from the first image are as close as possible as those of the second image, conserving the relative positioning of these key points but changing the perspective. Then key points from both images were superimposed on a plane and we computed the Euclidean distance between all pairs of key points in a pair of photographs, hence obtaining our SIFT-based distance. We used the implementation of SIFT and homography in the open source `openCV` library [38] version 3.4.

Using deep metric learning and triplet loss with CNN

We trained a CNN model using the triplet loss [39], in line with recent studies on other species [17, 18]. The triplet loss principle (details in [18]) relies on triplets of images, each triplet composed by a first image called *anchor* and another *positive* image of the same class (same giraffe here) with a third *negative* image of another class (any different giraffe). The training step consists in optimising the CNN model such that the Euclidean distance (computed using the last CNN layer) between any anchor and its positive image is minimal, while maximizing the distance between this anchor image with its negative image. We used an improved algorithm called *semi-hard triplet loss* [40] consisting in dealing with triplets where the positive and negative images are close enough to be informative still during the training procedure, using the TripletSemiHardLoss function in TensorFlow Addons. After training completion, we computed the Euclidean distances between any pair of giraffe flank photographs, again using the vector composing the last layer of our CNN model.

The CNN requires a training data set that we derived from the photograph clusters identified by the SIFT algorithm. We retained only those clusters fulfilling the following conditions: (*i*) the cluster was made of at least five images; (*ii*) the cluster demonstrated a perfect and verified consistency. This second condition is of upmost importance because errors in the training data set would lead to sub-optimal performances of the machine learning approach. We therefore carefully checked, manually, that the SIFT-based clusters we retained were perfectly unambiguous. We achieved this high level of data quality by discarding all cases where two or more giraffes overlapped on the same frame, or when giraffes were indifferently oriented from the back to the front (orientation ambiguities). We performed transfer learning using the pre-trained model MobileNetV2 readily available in Keras (input images resized to 224 x 224 pixels). We ran the model training stage with 200 epochs with batches of size 96. Again, we augmented our data (rotation, width and height shift, brightness, zoom and shear mapping) using the stochastic gradient descent optimizer with a rate of 0.2. Training took about 2 hours to complete on a Titan X card. Our pipeline was implemented with Keras 2.3.0.

Image similarity network, community detection and clusters of flank images

Following the computation of distances between all pairs of giraffe flanks obtained either with the SIFT or the CNN approach, we searched for clusters of flank images that should come from one single individual giraffe. We first defined a network made of nodes and representing giraffe flank images, and of edges: we considered that two nodes were connected by an edge, *i.e.* two flanks

were similar and came from the same giraffe individual, if the Euclidean distance between paired images felt below a given threshold (see below for more details). Therefore, the so-called *connected components* of this network are supposed to gather images from different individuals.

We estimated the distance threshold value by taking advantage of a property of complex networks called the *explosive percolation* [41]. The explosive percolation predicts a phase transition of the network just above a threshold point. At this turning point, adding a small number of edges in the network, for example by slightly increasing the distance threshold [42], leads to the sudden appearance of a *giant component* encompassing the majority of nodes. In other words, a small increase of the distance threshold leads to considering that almost most of the images come from the same giraffe individual. The threshold value was found graphically and was estimated to 340 when dealing with our SIFT-based distance (see Figure A.1a) and between 0.6 and 0.8 in our repeated experiments using our CNN-based distance (see Figure A.1b).

An additional issue need to be resolved: a connected component might be composed of different sub-components that are erroneously connected. This is the case when the distance computation fail in avoiding false positive edges (example in Figure A.2), i.e. when two flanks are erroneously considered similar. Moreover, in some cases the body of two or more giraffes can overlap in one photograph. In this situation, two or more nodes might be linked by edges, when we actually have different giraffes. Therefore, we applied a network clustering algorithm called *community detection*, developed in network science [43]. It is designed to split – only when relevant – any connected component into different groups of nodes that are significantly much more connected between themselves than with the others. Indeed, the presence of many edges inside a group of images suggested it was consistent (i.e. from the same individual), whereas the absence of many edges between two groups clearly informed about their inconsistency and heterogeneity (i.e. from two different individuals). We applied the community detection with the InfoMap algorithm [44] that is based on a random walker exploring the network. The final product of the community detection algorithm is a series of *clusters* of images corresponding either to a connected component or to a community retrieved by InfoMap. Any cluster is considered as a group of images coming from a single individual, and should also be the only cluster dealing with this particular individual.

Training and evaluation of CNN-based re-identification

We evaluated the overall predictive performance of our CNN deep metric learning with a classical 5-fold cross-validation procedure. We first randomly extracted 20% of our individuals from our original data set. Every image of these individuals (hereafter named *unknown* individuals) was moved to the test set. We then partitioned the images of the remaining 80% of our individuals as follows. Since we wanted to check the model ability to identify unknown as well as already known individuals, we moved two images of the these individuals (hereafter named *known individuals*) to the test set. Therefore, the test set contained all the images of the unknown individuals, and two images of the known individuals. We built the training set with the remaining data.

Our CNN model was not designed to classify images directly (i.e. assigning images to individual giraffes). Indeed, deep metric learning led to the computation of the distance between any pair of images. To mimic re-identification *per se*, we considered that we had a "reference book" with three *representative* images per known individuals (images from the training set). A representative image was intended to work as a "hook" to catch images from the same individual inside the test set. Indeed, we expected a small distance between an image in the test and a representative one when they corresponded to the same known individual. Finally, we calculated the distance between any pairs of images (as described before) in the three categories: *i*) the unknown individuals and *ii*) the known individuals from the test set, and *iii*) the representative images per known individual

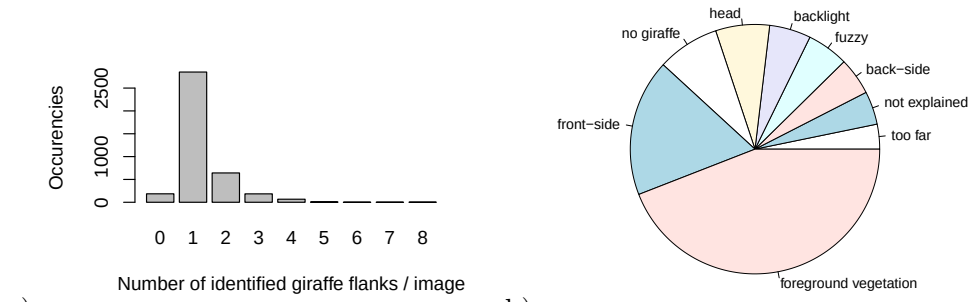
coming from the training set.

We quantified the predictive performance of the trained CNN model with three different indices measuring different facets of the performance. (*i*) The *top 5* accuracy consists in checking for each query image if another image from the same individual is among the five images with smallest distance. We also computed the widely used (*ii*) *sensitivity* and (*iii*) *specificity* for the evaluation of binary classification tests. Indeed, we built a network where the nodes are the giraffe flank images from the three categories (representative, known and unknown) and the edges are defined using a threshold (see previous section). Therefore, we were able to test the concordance between our "true" clusters (*i.e.* the validated SIFT-based clusters) and the "predicted" clusters obtained by applying the network-based approach on the CNN-based distance. Hereafter, we established the correspondence between true and predicted clusters by checking the predicted cluster with highest number of images from this true cluster. Sensitivity is the fraction of images of a true cluster correctly assigned, averaged over all the true clusters of interest. Specificity is the fraction of images in a predicted cluster that are in the corresponding true cluster, that we averaged over the true clusters.

Results

From thousands of images to hundreds of identified giraffe individuals

We trained the object detection method RetinaNet [26] on a set of 400 photographs for which the cropping of the giraffe flank has been previously done manually. When applying the automatic cropping procedure on our 3,940 photographs (see Figure 1a), we retrieved 5,019 images with associated bounding boxes, supposed to contain a single giraffe flank (see Figure 2a). The cropping failed for 186 photographs (failure rate: 4.7%), mostly due to foreground vegetation and unusual and difficult orientation of giraffes in the photograph (see examples on Figure 1b). In a few cases, a bounding box could contain the bodies of two overlapping giraffes, one being partially in front of the other (see Figure 2a). Similarly, we observed rare instances where a group of giraffes very close to each other is present on a photograph, RetinaNet could fail in retrieving the exact boundaries of each giraffe flank (see the worst case that we experienced, from a partially blurry photograph in Figure 2b).



a) b)

Figure 1: Performance of RetinaNet flank detection. a) Number of identified flanks per image. b) Manual classification of cropping problems encountered in 186 images where Retinanet failed to identify a giraffe flank.

a)  b) 

Figure 2: Examples of cropping with RetinaNet. a) Best-case scenario b) Worst-case but rare scenario where different individuals are overlapping and the right-end side of the photo is blurry.

Running the SIFT algorithm [8] to compare all pairs of flanks took about 800 CPU hours of heterogeneous computing resources. Our customised SIFT-based distance led to an image similarity network composed of 5,019 nodes and 11,249 edges, composed by 1,417 connected components (see Methods) among which 781 singletons. We identified several false positive edges leading to two images being erroneously classified as similar. False positive matches occurred either because of image background similarity (*e.g.* the same tree appeared on two images; see nodes 3 and 4 in Figure 3) or because of a perfect matching between giraffe orientation (see Figure A.2). These false positive matches connected images from different individuals inside a connected component. For instance, when the body of two giraffes overlapped on the same image (see node 2 in Figure 3), then this image linked two sets of images corresponding to the two individuals.

Our network-based approach, relying on community detection, dealt with these difficult cases to retrieve consistent *clusters* of flank images (different colors in Figure 3). The cluster size distribution is by definition more concentrated after network clustering (see Figure 4) with a maximal size of 35 instead of 373. Indeed, this very large connected component was clearly an artifact due to a chain of giraffe overlaps, and has been successfully split by our procedure (see Figure 5). We detected 316 clusters with more than 5 images, and 105 with more than 10 images. However, in rare cases, some images from the same individuals were found in different clusters (see Figure 5). Because these clusters arose from a single connected component, we could *a posteriori* check for consistencies by comparing clusters of the same component manually (such as performed for Figure 5).

From identified giraffe individuals to a deep learning approach for re-identification

We saved 178 human-validated unambiguous SIFT-based clusters to generate a reference data set of unique individual giraffes. Those 178 clusters were made of 1,393 images of giraffe flanks from which we evaluated our re-identification pipeline based on deep metric learning. Top 5 accuracy was about 90% on average (see Table 1) but note that the top 5 accuracy is of similar magnitude between both categories of known and unknown individuals. Top 5 accuracy increased, however, to 95.4% when considering only images of nknown individuals. Regarding the mapping between original and predicted clusters of images, both sensitivity and specificity were high, reaching values $> 90\%$ when dealing with known individuals (see Figure 6a). However, sensitivity dropped down to 53.9% when dealing with unknown individuals. This was partly due to the number of singletons image that are not associated to any other image in our network-based approach. Interestingly, specificity was moderate with all images included (about 71%) but increased to 90% dealing with only the images of unknown individuals (see Figure 6b).

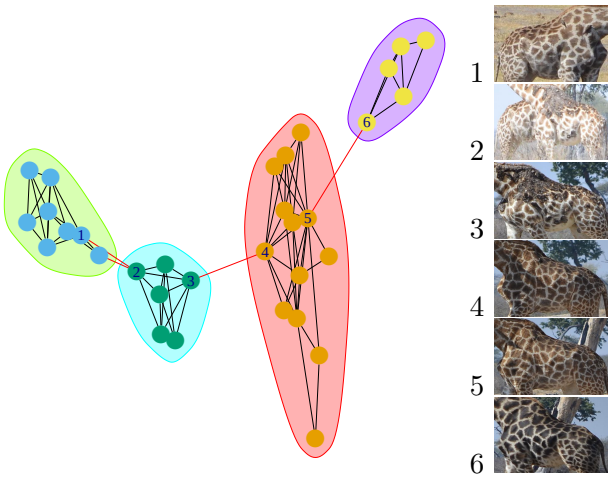


Figure 3: Example of a connected component split into four clusters using the InfoMap algorithm (see Methods). Each cluster is delineated by an ellipse of different color. Node 2 is an image with two giraffes that we also have in images 1 and 3 respectively. Images 3 and 4 are considered similar because of the presence of the same tree in the background (same for images 5 and 6).

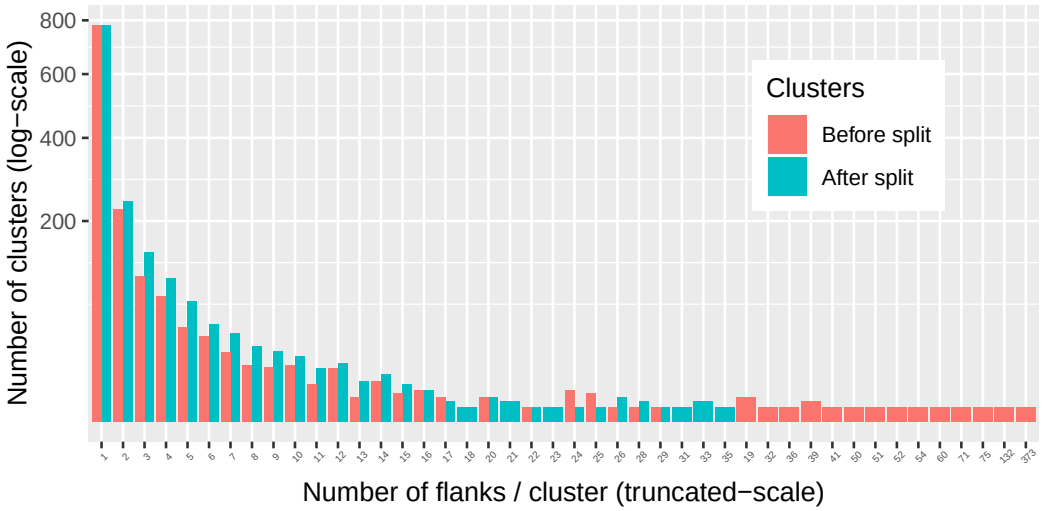


Figure 4: Re-identification from 5,019 giraffe flank images. Number of flank images retrieved by clusters, with the original clusters/connected component (red) or with the clusters retrieved using the InfoMap algorithm to split the connected components (blue; see Methods).

Discussion

In this study, we were able to propose a complete and fully automated pipeline to build a data set of re-identified giraffe individuals. We took advantage of the most recent techniques to perform object detection and crop the giraffe flanks to allow coat pattern analysis. This step was particularly efficient when giraffe individuals were not overlapping in a photograph. However, cascade of problems arose when overlap existed, including erroneous cropping and difficulties to assign a bounding box to a single individual (since it contained two individuals). We then followed previ-

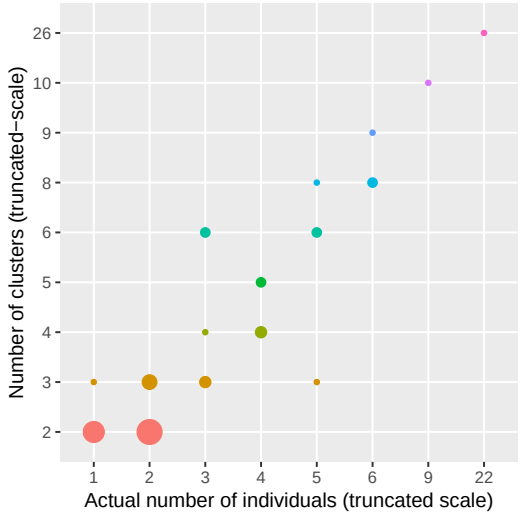


Figure 5: Agreement between the number of clusters (when at least two clusters were found out of a connected component) as returned by our machine-learning approach, and the human-based and manually-checked number of individuals. Circle size is proportional to the number of observations.

	Top 5 (%)	Sens. (%)	Spec. (%)	Singleton (%)
Known indiv.	89.4 ± 0.5	90.2 ± 0.4	90.8 ± 0.3	15.2 ± 1.1
Unknown indiv.	89.8 ± 0.7	53.9 ± 1	71.3 ± 1.8	23.3 ± 1.8
Unknown indiv. only	95.4 ± 0.4	53.8 ± 1	90.6 ± 1.1	23.3 ± 1.8

Table 1: Metric learning performance computed on the test set with images of known individuals and unknown individuals, or solely on a restricted test set with only the unknown individuals (excluding images from the known individuals). Top 5 accuracy, average sensitivity and specificity. Mean ± standard error over 10 trials.

ous attempts to use the SIFT-operator to perform pattern matching between flanks. However, we proposed a new problem statement using a network-based approach that efficiently handled false positive problems (*e.g.* matching between two images due to a similar element in the background) and false negative problems (e.g. differences in lighting and orientation). We were then able to go one step further other approaches from the literature since the end product of our automatic method is a comprehensive list of clusters of images, one cluster per re-identified individual.

We then evaluated the possibility to train a convolutional network to allow for giraffe identification when known and unknown individuals are present in a data set to be analysed. We observed that the CNN model seemed to contain relevant features that can be generalised to unknown individuals. One advantage of this approach is that our CNN-based distances were computed in a few minutes instead of hours to compute the SIFT-operator (once the training has been performed; see Table A.1). We achieved about 90% top 5 accuracy, which can definitely help field researchers in identifying giraffes, but is not fully satisfying if we are interested in a fully end-to-end automated pipeline that would require no checking by people. For this reason, we considered the same network-based approach using the prediction performed by the CNN. Whereas we obtained good results dealing with known individuals (i.e. good match between true and predicted cluster), we were not able to handle correctly the prediction of clusters of unknown giraffe images. We obtained

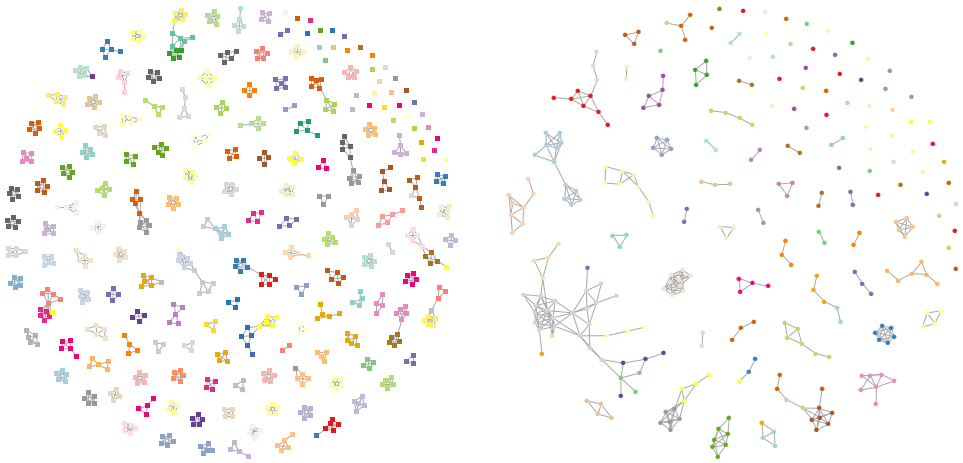


Figure 6: Image similarity networks defined by CNN-based metric learning. Node color represents different individual giraffes. Left: between flanks from known individuals only (including representative flanks; see Methods). Right: between flanks from unknown individuals only.

limited specificity due to many erroneous false positive links between images of known and unknown individuals. Even worse, we faced a sensitivity issue with a lot of images that were not connected to any other images. We understood that our CNN is vulnerable to orientation variation, much more than the more robust SIFT operator.

Further methodological developments should be investigated, to improve the current results. Firstly, our study suggests that a two step procedure could be developed: the first step consisting in detecting (and putting aside) images of known individuals; the second step consisting in treating the unknown individuals, as our study has shown that a CNN model can be trained to contain features that can be generalised to unknown giraffe individuals. Secondly, we believe that different data augmentation strategies should be able to improve our results, in particular in reducing the sensitivity of the CNN-based approach to image orientation changes. Such methodological developments will be able to benefit from this study and the large ground-truth data set that we make available upon request. In any case, we believe that our study, fully and freely reproducible, paves the way for future CNN-based works on animal individual re-identification.

Finally, this inter-disciplinary work provides guidelines about best practices to collect identification images in the field, if to be used later with an automated pipeline such as the one presented here. Better results can be achieved with simple framing rules of animals with cameras. First the field operator should try to avoid as much as possible overlaying bodies of two or more individuals as this was the most acute issue in our giraffe experience. Note that several but well separated individuals in the same photograph is not a problem at all thanks to the CNN cropping performed as a preliminary stage. Another point to pay attention to is the background which, if too similar on the same images (*e.g.* photographs shot from the very same spot) with obvious structures (tree, pond, rocks...) will likely mislead the computer vision algorithm, even on cropped images because cropping is rectangular and do not delineate the animal body. This situation often arises while photographing animals moving in line, as giraffes and many others often do. A last point is the heterogeneity of situations under which animals were observed. We did our best to improve the training data set with data augmentation. However, photographing giraffes in as many different conditions as possible would most likely improve the results. This includes light conditions (dawn,

dusk, noon), orientation of individual or background (open vs. more densely vegetated areas). More specific to CNN re-identification is the need to have a greater number of pictures (> 50) of photographs per individuals that what is currently available, so a particular attention should be given, in the field under optimal shooting conditions, to the opportunity to take more photographs of each observed individual.

Acknowledgments

We would like to thank Jeanne Duhayer for her considerable help in analysing our preliminary findings, and our statistician colleagues Laurent Jacob and Franck Picard for their insights on deep learning. This work was performed using the computing facilities of the CC LBBE/PRABI. Funding was provided by the French National Center for Scientific Research (CNRS) and the Statistical Ecology Research Group (EcoStat) of the CNRS. We are also grateful to Derek Lee for his kind advice in processing photographs, and for sharing with us his experience in the monitoring of giraffes. Finally, we acknowledge the support from the CNRS Zone Atelier / LTSER program for fieldwork.

References

[1] Tim Clutton-Brock and Ben C Sheldon. Individuals and populations: the role of long-term, individual-based studies of animals in ecology and evolutionary biology. *Trends in ecology & evolution*, 25(10):562–573, 2010.

[2] Loren D Hayes and Carsten Schradin. Long-term field studies of mammals: what the short-term study cannot tell us. *Journal of Mammalogy*, 98(3):600–602, 2017.

[3] Richard Despard Estes. The behavior guide to african mammals: including hoofed mammals, carnivores. *Primates*, pages 509–519, 1991.

[4] Ben G Weinstein. A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3):533–545, 2018.

[5] Douglas T Bolger, Thomas A Morrison, Bennet Vance, Derek Lee, and Hany Farid. A computer-assisted system for photographic mark–recapture analysis. *Methods in Ecology and Evolution*, 3(5):813–822, 2012.

[6] Kelly M Halloran, James D Murdoch, and Matthew S Becker. Applying computer-aided photo-identification to messy datasets: a case study of t hornicroft's giraffe (g iraffa camelopardalis thornicrofti). *African Journal of Ecology*, 53(2):147–155, 2015.

[7] Patrick Buehler, Bill Carroll, Ashish Bhatia, Vivek Gupta, and Derek E Lee. An automated program to find animals and crop photographs for individual recognition. *Ecological informatics*, 50:191–196, 2019.

[8] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[9] DT Bolger, B Vance, TA Morrison, and H Farid. Wild id user guide: pattern extraction and matching software for computer-assisted photographic mark, 2011.

[10] Sylvain Christin, Éric Hervet, and Nicolas Lecomte. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10(10):1632–1644, 2019.

[11] Aakash Lamba, Phillip Cassey, Ramesh Raja Segaran, and Lian Pin Koh. Deep learning for environmental conservation. *Current Biology*, 29(19):R977–R982, 2019.

[12] Di Wu, Si-Jia Zheng, Xiao-Ping Zhang, Chang-An Yuan, Fei Cheng, Yang Zhao, Yong-Jun Lin, Zhong-Qiu Zhao, Yong-Li Jiang, and De-Shuang Huang. Deep learning-based methods for person re-identification: A comprehensive review. *Neurocomputing*, 2019.

[13] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[14] Matthias Körschens, Björn Barz, and Joachim Denzler. Towards automatic identification of elephants in the wild. *arXiv preprint arXiv:1812.04418*, 2018.

[15] Mark F Hansen, Melvyn L Smith, Lyndon N Smith, Michael G Salter, Emma M Baxter, Marianne Farish, and Bruce Grieve. Towards on-farm pig face recognition using convolutional neural networks. *Computers in Industry*, 98:145–152, 2018.

[16] Andre C Ferreira, Liliana R Silva, Francesco Renna, Hanja B Brandl, Julien P Renoult, Damien R Farine, Rita Covas, and Claire Doutrelant. Deep learning-based methods for individual recognition in small birds. *bioRxiv*, page 862557, 2019.

[17] Olga Moskvyak, Frederic Maire, Asia O. Armstrong, Feras Dayoub, and Mahsa Baktashmotlagh. Robust re-identification of manta rays from natural markings by learning pose invariant embeddings, 2019.

[18] Soren Bouma, Matthew D. M. Pawley, Krista Hupman, and Andrew Gilman. Individual common dolphin identification via metric embedding learning, 2019.

[19] Daniel Schofield, Arsha Nagrani, Andrew Zisserman, Misato Hayashi, Tetsuro Matsuzawa, Dora Biro, and Susana Carvalho. Chimpanzee face recognition from videos in the wild using deep learning. *Science advances*, 5(9):eaaw0736, 2019.

[20] Qi He, Qijun Zhao, Ning Liu, Peng Chen, Zhihe Zhang, and Rong Hou. Distinguishing individual red pandas from their faces. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 714–724. Springer, 2019.

[21] Robert Bogucki, Marek Cygan, Christin Brangwynne Khan, Maciej Klimek, Jan Kanty Milczek, and Marcin Mucha. Applying deep learning to right whale photo identification. *Conservation Biology*, 33(3):676–684, 2019.

[22] Stefan Schneider, Graham W Taylor, and Stefan C Kremer. Similarity learning networks for animal individual re-identification-beyond the capabilities of a human observer. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops*, pages 44–52, 2020.

[23] Peng Chen, Pranjal Swarup, Michal Matkowski Wojciech, Adams Wai Kin Kong, Su Han, Zhihe Zhang, and Hou Rong. A study on giant panda recognition based on images of a large proportion of captive pandas. *Ecology and Evolution*, 2020.

[24] Stefan Schneider, Graham W Taylor, Stefan Linquist, and Stefan C Kremer. Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, 10(4):461–470, 2019.

[25] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[27] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[28] Bo Wang, Armin Pourshafeie, Marinka Zitnik, Junjie Zhu, Carlos D Bustamante, Serafim Batzoglou, and Jure Leskovec. Network enhancement as a general method to denoise weighted biological networks. *Nature communications*, 9(1):1–8, 2018.

[29] Simon Chamaillé-Jammes, Marion Valeix, Mathieu Bourgarel, Felix Murindagomo, and Hervé Fritz. Seasonal density estimates of common large herbivores in hwange national park, zimbabwe. *African Journal of Ecology*, 47(4):804–808, 2009.

[30] Z Muller, F Bercovitch, R Brand, D Brown, M Brown, D Bolger, K Carter, F Deacon, JB Doherty, J Fennessy, S Fennessy, AA Hussein, D Lee, A Marais, M Strauss, A Tutchings, and T Wube. *Giraffa camelopardalis* (amended version of 2016 assessment). the IUCN Red List of threatened species 2018: e.t9194a136266699. `http://dx.doi.org/10.2305/IUCN.UK.2016-3.RLTS.T9194A136266699.en`, 2018. [Downloaded on 13 September 2019].

[31] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[32] Jason Parham, Charles Stewart, Jonathan Crall, Daniel Rubenstein, Jason Holmberg, and Tanya Berger-Wolf. An animal detection pipeline for identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1075–1083. IEEE, 2018.

[33] Stefan Schneider, Graham W Taylor, and Stefan Kremer. Deep learning object detection methods for ecological camera trap data. In *2018 15th Conference on Computer and Robot Vision (CRV)*, pages 321–328. IEEE, 2018.

[34] Mohammad Sadegh Norouzzadeh, Dan Morris, Sara Beery, Neel Joshi, Nebojsa Jojic, and Jeff Clune. A deep active learning system for species identification and counting in camera trap images. *arXiv preprint arXiv:1910.09716*, 2019.

[35] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[36] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 456–473, 2018.

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[38] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[39] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[40] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[41] Dimitris Achlioptas, Raissa M. D'Souza, and Joel Spencer. Explosive percolation in random networks. *Science*, 323(5920):1453–1455, 2009.

[42] Satoru Hayasaka. Explosive percolation in thresholded networks. *Physica A: Statistical Mechanics and its Applications*, 451:1–9, 2016.

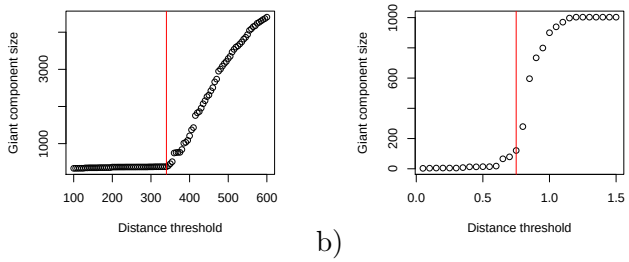
[43] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[44] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

Appendix

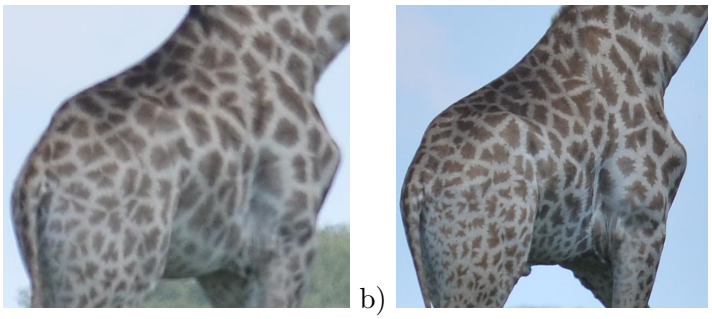
Task	Avg. computing time
SIFT-based distance	about 30 hours
CNN training (with GPU)	about 2 hours
CNN-based distance (with GPU)	2 minutes

Table A.1: Computing time to process 1003 images (training with 426 images only). Intel Xeon CPU E5-2650 v4 2.20GHz (CPU) and Nvidia Titan X card (GPU).



a) b)

Figure A.1: Giant component appearance. We manually estimated the threshold value (red line) used to build our image similarity network. The threshold is a) 340 when dealing with the SIFT-based distance and b) 0.75 in one of our repeated experiments using the CNN-based distance.



a) b)

Figure A.2: Rare SIFT false positive due to perfect shape and orientation matching. Two different giraffes have a similar pose in a) and b) and the SIFT-based distance between the two images is small and below the used threshold.

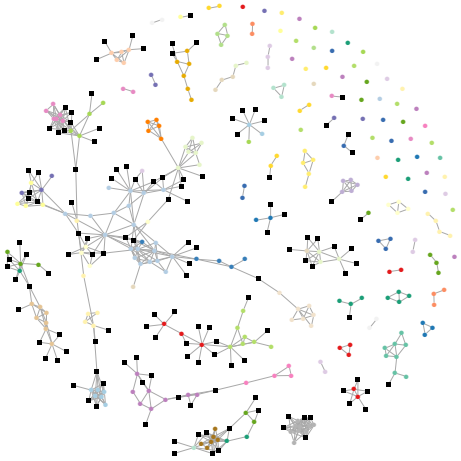


Figure A.3: Image similarity networks defined by CNN-based metric learning. Nodes'color correspond to giraffe individuals. Between flanks from unknown individuals and known individuals (black square). Edges between known individuals were discarded.