EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Automated identification of animal species in camera trap images

Xiaoyuan Yu[1,2], Jiangping Wang[2*], Roland Kays[3,4,5], Patrick A Jansen[3,6], Tianjiang Wang[1] and Thomas Huang[2]

Abstract

Image sensors are increasingly being used in biodiversity monitoring, with each study generating many thousands or millions of pictures. Efficiently identifying the species captured by each image is a critical challenge for the advancement of this field. Here, we present an automated species identification method for wildlife pictures captured by remote camera traps. Our process starts with images that are cropped out of the background. We then use improved sparse coding spatial pyramid matching (ScSPM), which extracts dense SIFT descriptor and cell-structured LBP (cLBP) as the local features, that generates global feature via weighted sparse coding and max pooling using multi-scale pyramid kernel, and classifies the images by a linear support vector machine algorithm. Weighted sparse coding is used to enforce both sparsity and locality of encoding in feature space. We tested the method on a dataset with over 7,000 camera trap images of 18 species from two different field cites, and achieved an average classification accuracy of 82%. Our analysis demonstrates that the combination of SIFT and cLBP can serve as a useful technique for animal species recognition in real, complex scenarios.

Keywords: Species identification; SIFT; cLBP; Feature learning; Max pooling; Weighted sparse coding

1 Introduction

Monitoring biodiversity, especially the effects of climate and land-use change on wild populations, is a critical challenge for our society [1]. Sensor networks are a promising approach for collecting the spatio-temporal data at scales needed to address this challenge [2], especially visual sensors that record images of animals that move across their field of view (i.e. camera traps [3,4]). However, processing the large volumes of images that such studies generate to identify the species of animals recorded remains a challenge.

At present, all camera-based studies of wildlife use a manual approach where researchers examine each photograph to identify the species in the frame. For studies collecting many tens or hundreds of thousands of photographs, this is a daunting task [5].

Computer-assisted species recognition on camera-trap images could make this work flow more efficient, and reduce, if not remove, the amount of manual work involved in the process. However, in comparison with the typical video from surveillance of building and street views, camera trap of animals amidst vegetation are more difficult to incorporate into image analysis routines because of low frame rates, background clutter, poor illumination, serious occlusion, and complex pose of the animals.

Inspired by recent object recognition works [6,7] in the computer vision community, we improved sparse coding spatial pyramid matching (ScSPM) method for species recognition on images collected by camera traps. During the local feature extraction, we combined dense scale-invariant feature transform (SIFT) [8] of features with cell structured local binary patterns (cLBP) [9] to represent the object of interest. We apply weighted sparse coding for dictionary learning, and thus enforce both sparsity and locality, since locality may be more important than sparsity, as suggested by Wang et al. [7]. Then we used linear SVM to classify image of species.

We tested our method with images collected by camera traps that were deployed in two different environments, tropical rainforest and temperate forest, that represent a wide variety of backgrounds and conditions. From this

*Correspondence: jwang63@illinois.edu
[2]Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL, USA
Full list of author information is available at the end of the article

collection, we selected sequences and species to keep the data balanced. Then, we manually cropped animals from all the frames to generate a dataset with 7,196 images over 18 different vertebrate species.

2 Related work

Most related works are camera-based studies of wildlife that use image analysis to identify individual animals of select species with unique coat patterns (e.g., spots or stripes). Bolger et al. [10] applied software to help identify individual animals based on coat patterns for subsequent photographic mark-recapture analysis. The data they used was image based, which is a cost-effective, non-invasive way to study population. The method they used was the SIFT key points extraction and matching. Thus, they only focused on individual animal identification for these strongly marked texture species.

Identifying species from remote camera images remains a major challenge that has not been addressed. In the community of computer vision, there exist a lot of methods to recognize general object. One of the most successful ones is Yang's work [6], in which ScSPM is applied. Spatial pyramid matching (SPM) with max pooling [11] can not only model the spatial layout of local image features, but also achieve translation invariance of animal body. As being easy and simple to construct, the SPM kernel turns out to be highly effective in practice [12]. Sparse coding has been successfully applied to model local features, and to construct overcomplete dictionary that can sparsely represent the local features. Sparse coding can yield better results than vector quantization and hard assignment [6].

3 Materials and methods

Our pattern extraction and classification program is based on the ScSPM [6], as shown in Figure 1. The algorithm first extracts local feature descriptor densely. We combine two kinds of local descriptors: SIFT and cLBP. In order to sparsely represent local features, the dictionary is learned via weighted sparse coding, for each kind of descriptor feature. Similar local features can generate similar codes after sparse coding on the dictionary, which is essential for recognition because it retains discriminative information while suppressing the noise. Finally, max pooling using SPM is used to construct the global image feature that converts an image or a bounding box to a single vector. We then apply linear multi-class SVMs to classify the global feature to one category of species, assuming SVMs are trained beforehand using training data.

3.1 Local feature extraction

The camera-trap images contain rich noise and clutter. This requires us to develop a both discriminant and invariant local feature to describe local image patches. Dense SIFT feature, also known as dense histogram of oriented gradients, is successfully used in some recognition work. SIFT descriptor is invariant to moderate scaling and shifting change of edges and linear illuminance variation in image patch; however, it fails when nonlinear illuminance change occurs. cLBP, in contrast, is the perfect local texture descriptor that is invariant to moderate nonlinear illuminance variation. In the area of computer vision, for human detection [13], HOG and cLBP features are concatenated to obtain the final feature. But the simple

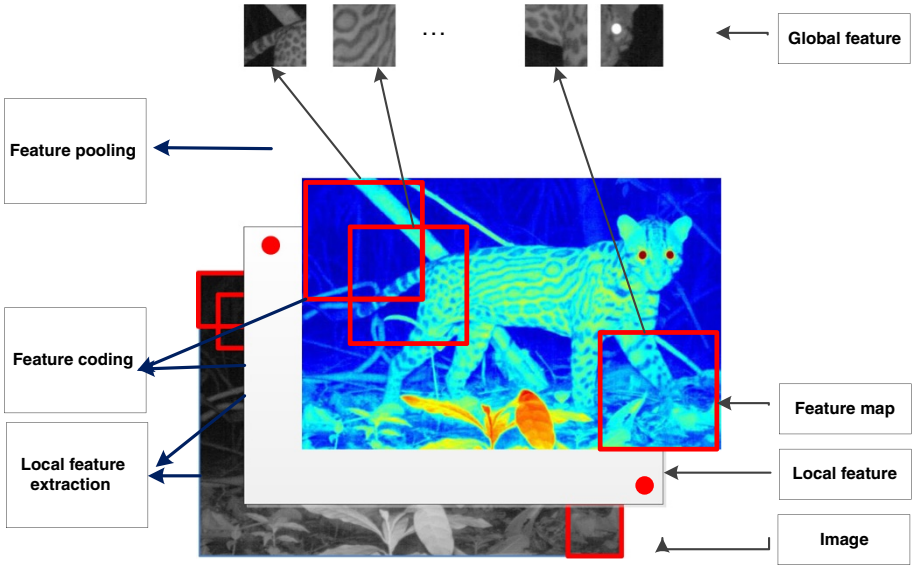

Figure 1 The architecture of ScSPM algorithm. The densely extracted local features are pooled across different spatial locations over different spatial scales.

concatenation would potentially cause the following problem: the feature space becomes more complex and more difficult to classify. We thus used the procedure of Zhang et al. [14] to extract HOG and cLBP, and concatenate responses only after coding them separately.

The SIFT descriptor is similar to the HOG. Both are histograms of oriented gradients. The SIFT descriptor is illustrated in Figure 2. After calculating the gradient map for each image, SIFT creates oriented gradient histograms for 4×4 grid regions, instead of 2×2 as in HOG. The full 128 dimensional SIFT descriptor is created by concatenating the 16 histograms in 16×16 image patch.

cLBP is a very good texture descriptor that extracts histogram of the LBP patterns from local cells, as shown in Figure 2. In order to filter out noises, LBP is modified into a uniform LBP pattern [15]. We use the notation $LBP^{u}_{n,r}$ to denote LBP feature that takes n sample point with radius r, and the number of 0-to-1 transitions is no more than u. The pattern that satisfies this constraint is called uniform pattern [15]. For example, the pattern 0010010 is a nonuniform pattern for LBP^2, and is a uniform pattern for LBP^4 because LBP^4 allows four 0-to-1 transitions. In our approach, we set $u = 2$, $n = 8$, and $r = 1$. In this setting, the dimension of LBP is 59.

The rationale for combination of SIFT and cLBP is that at pixel level, the oriented gradient has been assigned to 8 bins in SIFT, while in uniform $LBP^2_{8,1}$ the number of bins is 59. At cell level, 16 cells are used in SIFT while only 1 cell is used in cLBP. So SIFT is very accurate at the cell level but invariant at the pixel level, while the opposite holds for cLBP. The combination of the two solves the trade-off between discrimination and invariance, at both the pixel and the cell level.

3.2 Dictionary learning and weighted sparse coding

The goal of dictionary learning is to capture high-level information, that is, to select some items to describe the distribution of the input space. We get a local image feature set $\boldsymbol{X}$ by randomly sampling in feature space. Then $\boldsymbol{X}$ approximates the distribution of the input space. But $\boldsymbol{X}$ contains a huge number of signals, which make it impossible to use $\boldsymbol{X}$ directly in coding. Dictionary learning aims to generate a compact dictionary that can sparsely represent the incoming signal with minimum error.

Let $\boldsymbol{X}$ be in a D-dimensional features space, i.e. $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$. The dictionary is $\boldsymbol{V} = [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K] \in \mathbb{R}^{D \times K}$ with K atoms. The traditional dictionary leaning and sparse coding method formulate the problem as follows:

$$\begin{aligned} &\min_{\boldsymbol{V},\boldsymbol{U}} \ \|\boldsymbol{X} - \boldsymbol{V}\boldsymbol{U}\|_2 + \lambda \|\boldsymbol{U}\|_1 \\ &\text{s.t.} \ \|\boldsymbol{v}_k\| \leq 1, \ \forall k = 1, 2, \cdots, K, \end{aligned} \tag{1}$$

where $\boldsymbol{U} = [u_1, \cdots, u_N] \in \mathbb{R}^{K \times N}$ is the matrix of sparse codes.

Inspired by the work of Wang et al. [7] in which encoding of features is based on the locality in the feature space, we adapt the original sparse coding to the weighted sparse coding as follows to enforce both sparsity and locality:

$$\begin{aligned} &\min_{\boldsymbol{V},\boldsymbol{U}} \ \|\boldsymbol{X} - \boldsymbol{V}\boldsymbol{U}\|_2 + \lambda \|W\boldsymbol{U}\|_1 \\ &\text{s.t.} \ \|\boldsymbol{v}_k\| \leq 1, \ \forall k = 1, 2, \cdots, K, \end{aligned} \tag{2}$$

where W is a diagonal weighting matrix whose elements are computed as

$$W_i(k,k) = \|\boldsymbol{X}_i - \boldsymbol{V}_k\|_2, k = 1, 2, \cdots, K. \tag{3}$$

Many algorithms have been proposed to solve this dictionary learning problem, e.g., [16]. $\boldsymbol{V}$ is well known as a codebook and can be trained and fixed in the testing phase. Recently, there has been a lot of work on supervised dictionary learning (e.g., [17,18]) to adapt the dictionary for classification purpose, but it is often computationally expensive and cannot handle large multi-class problem well. Thus, our work employs unsupervised dictionary learning using weighted sparse coding, as in Equation 2.

3.3 Linear SPM and multi-scale max pooling

Spatial pyramid matching is an extension of Bag of Words (BoW) method, and it models the spatial layout of local image features at multiple scales. Figure 1 illustrates the

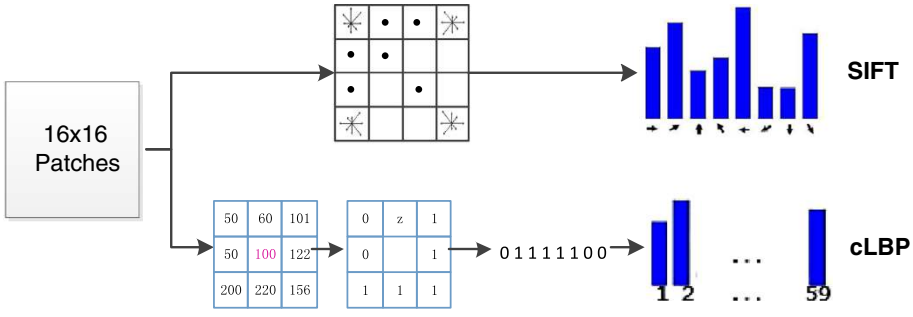

Figure 2 The procedure of extracting local features. First, calculate the gradients and LBP patterns on raw pixel patch. Second, create histogram features for SIFT and LBP, respectively .

Table 1 The numbers of sequences for each species

Common name	Number of total seq	Number of remained seq	Remained frames
Agouti	5,423	100	950
Collared peccary	904	100	901
Paca	298	100	1,196
Red brocket deer	588	100	982
White nosed coati	203	100	1,313
Spiny rat	130	100	712
Ocelot	345	100	548
Mouflon	216	100	2,365
Red deer	462	100	2,830
Wild boar	240	100	1,883
Wood mouse	264	100	1,350
Red squirrel	160	99	645
Great tinamou	130	99	1,194
Roe deer	620	99	1,271
Common opossum	93	93	916
White-tailed deer	93	92	2,226
European hare	87	87	700
Red fox	70	70	551
European badger	42	0	0
Panamanian white-throated capuchin	42	0	0
Northern tamandua	36	0	0
Brown four-eyed opossum	29	0	0
Coiba Island white-tailed deer	28	0	0
Nine-banded long-nosed armadillo	24	0	0
Tayra	24	0	0
Common raven	19	0	0
Gray four-eyed opossum	10	0	0
Ruddy quail dove	9	0	0
Eurasian jay	9	0	0
European pine marten	9	0	0
Blackbird	8	0	0
Baird's tapir	7	0	0
Crested guan	6	0	0
Forest rabbit	6	0	0
Unknown dutch mouse	6	0	0
White-faced capuchin	5	0	0
Skylark	5	0	0
Great spotted woodpecker	5	0	0
Turkey vulture	4	0	0
Spotted antbird	3	0	0
Howler monkey	3	0	0
Green iguana	3	0	0
Song thrush	3	0	0
Wood pigeon	3	0	0
Carrion crow	3	0	0
Common buzzard	3	0	0
Robinson's mouse opossum	2	0	0
Bank Vole	2	0	0
Black eared opossum	1	0	0
Crab eating racoon	1	0	0
Lizard	1	0	0
Armadillo	1	0	0
European rabbit	1	0	0
Chaffinch	1	0	0
Goshawk	1	0	0
Blue tit	1	0	0
European robin	1	0	0

whole structure of ScSPM. Let $\boldsymbol{U}$ be the matrix of sparse codes of applying Equation 2 to a descriptor set $\boldsymbol{X}$, assuming the codebook $\boldsymbol{V}$ is pre-computed. The pooled features from various locations and scales are then concatenated to form a spatial pyramid representation of the image. In each pyramid, a max pooling function is applied on the absolute sparse codes:

$$z_j = \max\{|u_{j1}|, |u_{j2}|, \cdots, |u_{jM}|\} \tag{4}$$

where z_j is the jth element of $\boldsymbol{z}$, u_{ji} is the matrix element at jth row and ith column of $\boldsymbol{U}$. Max pooling is beneficial for translation invariance because the maximum response will be filtered out if it is a small translation.

Let image I_i be represented by $\boldsymbol{z}_i$, a simple linear SPM kernel is defined by [6]

$$\kappa(\boldsymbol{z}_i, \boldsymbol{z}_j) = \boldsymbol{z}_i^T \boldsymbol{z}_j \tag{5}$$

With linear SPM kernel, we can directly use linear SVM, for which the training cost is $O(n)$ in computation, and the testing cost for each image depends on the dimension of feature.

3.4 Multi-class linear SVM

Let $\{(\boldsymbol{z}_i, y_i)\}_{i=1}^n, y_i \in \mathcal{Y} = \{1, 2, \cdots, L\}$ be the training data. We stick to the implementation in Yang et al. [6], and use one-against-all strategy to train L binary linear SVMs that each solve the following unconstrained convex optimization problem:

$$\min_{w_c} J(w_c) = \|w_c\|^2 + C \sum_{i=1}^{n} l(w_c; y_i^c, \boldsymbol{z}_i), \tag{6}$$

Figure 3 Two sequences of agouti and collared peccary captured in day and night.

Table 2 The 18 terrestrial species, captured by camera traps in Panama and the Netherlands

Common name	Latin name	Pictures (*n*)	Site
Agouti	*Dasyprocta punctata*	518	Panama
Paca	*Cuniculus paca*	285	Panama
Collared peccary	*DTayassu tajacu*	263	Panama
Red brocket deer	*Mazama americana*	297	Panama
White-nosed coati	*Nasua narica*	325	Panama
Spiny rat	*Proechimys semispinosus*	175	Panama
Ocelot	*Leopardus pardalis*	184	Panama
Red-tailed squirrel	*Sciurus granatensis*	143	Panama
Common opossum	*Didelphis marsupialis*	264	Panama
Great tinamou	*Tinamus major*	350	Panama
White-tailed deer	*Odocoileus virginianus*	1,091	Panama
Mouflon	*Apodemus sylvaticus*	896	Holland
Red deer	*Cervus elaphus*	802	Holland
Roe deer	*Capreolus capreolus*	362	Holland
Wild boar	*Sus scrofa*	487	Holland
Red fox	*Vulpes vulpes*	120	Holland
European hare	*Lepus europaeus*	176	Holland
Wood mouse	*Apodemus sylvaticus*	455	Holland

Images were used to test the recognition algorithm.

where $y_i^c = 1$ if $y_i = c$, otherwise $y_i^c = -1$, and $l(w_c; y_i^c, \boldsymbol{z}_i)$ is the hinge loss function. The standard hinge loss function is not differentiable everywhere, but here we can use quadratic hinge loss as below instead to make use of gradient-based optimization methods, e.g., LBFGS [6].

$$l(w_c; y_i^c, \boldsymbol{z}_i) = [\max(0, 1 - w_c^T \boldsymbol{z} \cdot y_i^c)]^2$$

4 Experimental results

4.1 Data set

We used images of wildlife captured with motion-sensitive camera traps (Reconyx RC55, PC800 and HC500, Holmen, WI, USA), which generate sequences of 3.1 Megapixel JPEG images at about 1 frame/s upon triggering by an infrared motion sensor. Color images are captured during the day and gray-scale images are captured at night using and an infrared flash, which is invisible to most animals. We used images from tropical rain forest (Barro Colorado Island, Panama) and temperate forest and heathland (Hoge Veluwe National Park, the Netherlands). Expert zoologists identified the animals in the images. We did not edit the data set for ease of identification, so it includes many of the typical challenges faced by camera trapping data, including cases where the animal is too small or is occluded by vegetation.

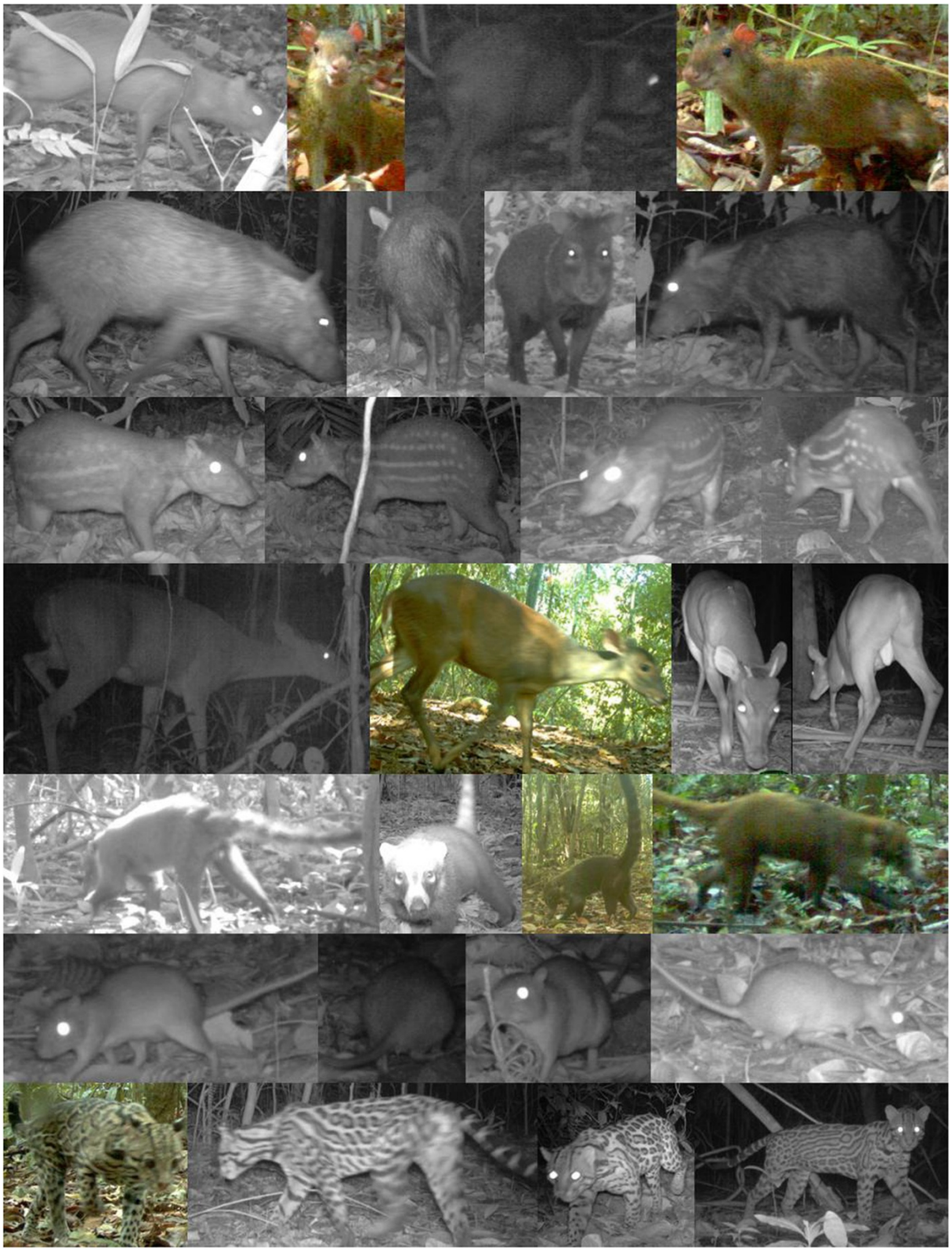

Figure 4 The cropped sample images. Each row contains a species. From top to bottom, they are the agouti, collared peccary, paca, red brocket deer, white-nosed coati, spiny rat, and ocelot. Each sample image has their own scale, aspect ratio, and pose.

Table 3 Confusion matrix of species recognition, obtained by averaging matrices resulting from 10 runs

	Agouti	Collared peccary	Paca	Red brocket deer	White-nosed coati	Spiny rat	Ocelot	Red squirrel	Common opossum	Great tinamou	White-tailed deer	Mouflon	Red deer	Roe deer	Wild boar	Red fox	European hare	Wood mouse
Agouti	*86.9*	1.6	0.4	0.1	1.2	1.0	0.0	0.8	0.5	3.7	1.9	0.1	0.3	0.4	0.6	0.0	0.3	0.0
Collared peccary	7.5	*77.1*	1.3	2.7	1.8	0.0	1.0	0.0	1.4	1.8	4.4	0.0	0.3	0.0	0.5	0.0	0.0	0.4
Paca	0.5	0.1	*90.7*	1.7	0.7	0.7	0.7	0.0	1.6	0.2	2.2	0.0	0.0	0.0	0.2	0.1	0.5	0.0
Red brocket deer	0.1	0.2	0.3	*58.1*	0.7	0.0	0.4	0.3	0.0	0.6	36.4	0.0	1.8	0.6	0.4	0.0	0.0	0.0
White-nosed coati	3.6	0.3	0.1	0.0	*88.5*	0.0	0.0	0.8	0.3	1.2	4.0	0.0	0.4	0.5	0.1	0.0	0.1	0.1
Spiny rat	2.0	0.2	0.4	0.0	0.0	*78.5*	0.0	1.3	6.9	1.9	1.3	0.9	0.0	0.0	0.6	0.9	1.9	3.3
Ocelot	0.0	0.4	1.1	0.2	0.2	0.0	*92.9*	0.0	0.7	0.0	2.9	0.0	0.5	0.0	0.2	0.4	0.7	0.0
Red squirrel	16.5	0.0	0.5	0.0	5.8	1.9	0.0	*64.7*	0.7	4.2	1.4	0.9	0.9	0.0	0.0	0.0	0.0	2.6
Common opossum	2.6	1.4	3.5	0.3	0.3	3.8	0.0	0.0	*79.1*	0.8	5.0	0.3	0.4	0.0	1.0	0.5	1.1	0.1
Great tinamou	6.3	0.2	0.3	0.3	1.3	0.0	0.0	0.0	1.9	*85.7*	2.4	0.7	0.5	0.1	0.3	0.0	0.0	0.0
White-tailed deer	0.6	1.2	0.4	2.9	0.8	0.1	0.0	0.1	0.0	0.5	*90.3*	0.4	1.2	0.2	0.5	0.0	0.2	0.6
Mouflon	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	*97.2*	1.3	0.2	0.0	0.1	0.4	0.0
Red deer	0.2	0.0	0.0	0.1	0.0	0.0	0.0	0.2	0.0	0.0	1.3	2.8	*92.7*	1.6	1.0	0.0	0.0	0.1
Roe deer	0.7	0.4	0.4	0.1	0.0	0.0	0.0	1.5	0.2	0.2	3.1	4.9	10.3	*76.4*	0.3	0.8	0.8	0.0
Wild boar	1.1	0.0	0.1	0.1	0.2	0.0	0.1	0.3	0.0	0.0	0.1	0.7	0.7	0.1	*96.4*	0.1	0.1	0.1
Red fox	1.9	0.0	0.3	0.0	0.0	0.6	0.0	0.3	3.3	0.6	3.9	14.2	6.7	5.8	5.0	*53.1*	4.4	0.0
European hare	0.9	0.2	0.9	0.0	0.0	0.8	0.2	0.9	0.2	0.0	2.6	12.1	1.7	4.2	2.5	3.2	*67.7*	1.9
Wood mouse	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	*100*

For the 18 species, accuracy averaged 82% with standard deviation of 0.9%.

Table 4 The confusion matrix of species recognition on Panama data

	Agouti	Collared peccary	Paca	Red brocket deer	White-nosed coati	Spiny rat	Ocelot	Red squirrel	Common opossum	Great tinamou	White-tailed deer
Agouti	90.1	1.2	0.6	0.2	0.8	0.6	0.0	0.6	1.3	2.9	1.7
Collared peccary	8.7	79.5	0.6	1.4	1.3	0.0	1.1	0.0	1.9	1.3	4.2
Paca	0.0	0.3	91.3	1.6	0.1	0.6	0.8	0.0	2.1	0.8	2.3
Red brocket deer	0.6	0.2	0.3	59.7	0.9	0.0	0.1	0.0	0.4	0.6	37.2
White-nosed coati	3.2	0.2	0.0	0.1	89.0	0.0	0.3	0.7	0.9	1.2	4.4
Spiny rat	2.2	0.6	1.1	0.0	0.0	84.3	0.0	0.4	8.1	2.0	1.3
Ocelot	0.0	0.2	3.4	0.2	0.4	0.2	91.4	0.0	0.2	0.0	4.1
Red squirrel	18.4	0.0	0.0	0.0	3.7	1.6	0.0	71.2	0.0	3.0	2.1
Common opossum	2.0	1.6	5.1	0.5	0.3	3.6	0.0	0.1	83.1	0.8	2.9
Great tinamou	4.5	0.0	0.0	0.0	1.0	0.2	0.0	0.2	1.5	90.6	1.9
White-tailed deer	1.0	1.1	0.4	3.2	0.5	0.2	0.1	0.1	0.3	0.7	92.4

For the 11 species, accuracy averaged 83.8% with standard deviation of 1.2%.

In total, we got 10,598 sequences over 57 species. The numbers of sequences of each species were unbalanced. As shown in Table 1, 40 out of 57 species have less than 50 sequences. we exclude these species and remain top 18 species. In order to build a balanced test data set, we chose up to 100 sequences from each species. Where the available number of sequences for a species was less than 100, we choose all of the sequences for that species. After such operation, 1,739 sequences for 18 species remained. Table 1 lists the number of remained sequences and frames for each species.

The camera trapped sequences are of low frame rate (1 frame/s) and short length (about 10 frames/sequence). Two typical image sequences are shown in Figure 3. The first two rows show consecutive frames of the agouti, in which the leaves dangled in the wind. The second two rows are continual frames of the collared peccary. If the peccary suddenly moved close to the camera, the illumination changes a lot because it cut out much of the light. The common motion detection method cannot handle this case very well. In order to get clear data, we manually cropped all the animals from the sequences. Since most of them are empty frames, in which the cameras are activated by motion from background, only 7,196 animal images are kept. Table 2 lists the details of the proposed dataset. During the progress of cropping, we kept the original animal size, color, and aspect ratio. Figure 4 shows the cropped samples for seven species.

4.2 Implementation and result

We developed a species recognition algorithm based on ScSPM, implemented as follows. The images were all converted into gray scale and both the SIFT descriptor and the cLBP descriptor were then extracted from 16×16 pixel patches. All the patches of each image were densely sampled on a grid with stepsize of 4 pixels. Both SIFT and cLBP were normalized to be unit norm with dimensions 128 and 59, respectively. For the dictionary learning process, we extracted SIFT and cLBP from 20,000 patches that are randomly sampled on training set. Dictionaries were trained for SIFT and cLBP separately, with the same dictionary size $K = 1,024$.

Following the standard benchmark procedures, we repeated the experimental process by 10 runs to obtain reliable results. In each run, we randomly selected 70% of the images of each species for training and kept the remaining 30% for testing. We report our final results as a confusion matrix.

We first test our approach on all 18 species, and the classification result is shown in Table 3. In real world scenarios, it is not necessary to distinguish species across the two-place datasets. Thus, we also test our method on the two datasets (Panama and Netherlands) separately. The classification results are shown in Tables 4 and 5.

Table 5 The confusion matrix of species recognition on Holland data

	Mouflon	Red deer	Roe deer	Wild boar	Red fox	European hare	Wood mouse
Mouflon	97.3	1.5	0.4	0.2	0.2	0.4	0.0
Red deer	2.7	93.6	2.0	1.4	0.0	0.2	0.2
Roe deer	6.3	9.6	81.5	0.7	0.5	1.3	0.1
Wild boar	0.5	1.0	0.1	98.0	0.0	0.1	0.2
Red fox	14.2	7.8	7.5	6.7	53.3	10.6	0.0
European hare	8.9	3.0	7.4	2.8	2.8	73.8	1.3
Wood mouse	0.0	0.0	0.0	0.0	0.0	0.0	100.0

For the 7 species, accuracy averaged 85.4% with standard deviation of 1.5%.

Since the SIFT and cLBP can describe the texture at different level, we did the experiment using SIFT, cLBP, and the combination of SIFT and cLBP, respectively, to show how the combination improved the performance. The SIFT feature is good at extracting the silhouette of an animal, while cLBP is powerful in describing the skin texture of animals. Thus, it is reasonable to combine SIFT and cLBP. As we can see in Table 6, SIFT feature is more discriminative than cLBP, and the performance is boosted much by combining them.

In Table 3, we can see that the overall accuracy is about 82%. Wood mouse is correctly recognized 100%, which is surprising, considering that none biometric features are used. For over one third of the 18 species, this experiment obtained classification accuracy over 90%, such as paca, ocelot, red deer, and wild boar. As expected, red brocket deer is easily misclassified as white-tailed deer because they are of the same ontology and have the similar appearance. In order to better classify the two species like these, biometric features, such as spots on the fur and shape of antlers, play a key role in species recognition. However, automatically identifying biometric features is a challenging task, to our best knowledge.

5 Conclusion

We have shown that object recognition techniques from computer vision science can be effectively used to recognize and identify wild mammals on sequences of photographs taken by camera traps in nature, which are

Table 6 Performance of different procedures for recognition of local image features

Feature	Average accuracy (%)	Standard deviation (%)
SIFT	78.9	0.7
cLBP	74.5	1.1
SIFT + cLBP	82.0	0.9

The combination of SIFT and cLBP improves performance a lot.

notorious for high levels of noise and clutter. Although some species are of the same ontology, the proposed method can detect imperceptible differences between them. The combination of SIFT and cLBP as descriptors of local images features significantly improved the recognition performance, which is abundant in texture description at multiple scales.

In the future work, some biometric features that are important for species analysis will be included in the local features, such as color, spots, and size of the body. Since the original sequences captured with motion-sensitive camera traps have motion information, we will develop an automatic animal segmentation algorithm in the future.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported in part by the National Science Foundation Grant DBI 10-62351. Field data were collected with support from the National Science Foundation (NSF-DEB 0717071 to R.W.K.) and the Netherlands Organization for Scientific Research (863-07-008 to P.A.J.). XY and TW would like to acknowledge support by the National Natural Science Foundation of China Grant 61073094.

Author details
[1]Department of Computer Science, Huazhong University of Science and Technology, Wuhan, Hubei, China. [2]Beckman Institute, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [3]Smithsonian Tropical Research Institute (STRI), Balboa, Ancon Panama, Republic of Panama. [4]North Carolina Museum of Natural Sciences, Raleigh, NC, USA. [5]Fisheries, Wildlife & Conservation Program, North Carolina State University, Raleigh, NC, USA. [6]Department of Environmental Sciences, Wageningen University, Wageningen, the Netherlands.

Received: 1 February 2013 Accepted: 21 August 2013
Published: 4 September 2013

References

1. Committee on Grand Challenges in Environmental Sciences NRCUC, *Grand Challenges in Environmental Sciences* (National Academies Press, Washingthon, DC, 2001)
2. J Porter, P Arzberger, H Braun, P Bryant, S Gage, T Hansen, P Hanson, C Lin, F Lin, T Kratz, T Williams, S Shapiro, H King, W Michener, Wireless sensor networks for ecology. BioScience. **55**(7), 561–572 (2005)
3. R Kays, S Tilak, B Kranstauber, P Jansen, C Carbone, M Rowcliffe, T Fountain, J Eggert, Z He, Monitoring wild animal communities with arrays of motion sensitive camera traps. Int J Res Rev Wireless Sensor Netw. **1**, 19–29 (2011)
4. J Aguzzi, C Costa, Y Fujiwara, R Iwase, Ramirez-E Llorda, P Menesatti, A novel morphometry-based protocol of automated video-image analysis for species recognition and activity rhythms monitoring in deep-sea fauna. Sensors. **9**(11), 8438–8455 (2009)
5. E Fegraus, K Lin, J Ahumada, C Baru, S Chandra, C Youn, Data acquisition and management software for camera trap data: a case study from the TEAM Network. Ecol. Inform. **6**(6), 345–353 (2011)
6. J Yang, K Yu, Y Gong, T Huang, Linear spatial pyramid matching using sparse coding for image classification, in *IEEE Conference on Computer Vision and Pattern Recognition*, (Miami, 20-25 June 2009), pp. 1794–1801
7. J Wang, J Yang, K Yu, F Lv, T Huang, Y Gong, Locality-constrained linear coding for image classification, in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (San Francisco, CA, 13-18 June 2010), pp. 3360–3367
8. D Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
9. T Ahonen, A Hadid, M Pietikainen, Face description with local binary patterns: application to face recognition. Pattern Anal. Mach. Intell, IEEE Trans. **28**(12), 2037–2041 (2006)
10. B Bolger, DT Morrison, TA Vance, D Lee, H Farid, A computer-assisted system for photographic mark–recapture analysis. Methods Ecol. Evol. **3**(5), 813–822 (2012)
11. T Serre, L Wolf, T Poggio, Object recognition with features inspired by visual cortex, in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, (San Diego, CA, 20-26 June 2005), pp. 994–1000
12. S Lazebnik, C Schmid, J Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, (New York, 17-22 June 2006), pp. 2169–2178
13. X Wang, T Han, S Yan, An HOG-LBP human detector with partial occlusion handling, in *2009 IEEE 12th International Conference on Computer Vision*, (Kyoto, Japan, 27 September - 4 October, 2009), pp. 32–39
14. J Zhang, K Huang, Y Yu, T Tan, Boosted local structured HOG-LBP for object localization, in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Colorado Springs, Colorado, 20-25 June 2011), pp. 1393–1400
15. T Ojala, M Pietikäinen, D Harwood, A comparative study of texture measures with classification based on featured distributions. Pattern Recognit. **29**, 51–59 (1996)
16. H Lee, A Battle, R Raina, A Ng, Efficient sparse coding algorithms. Adv. Neural Inf. Process. Syst. **19**, 801 (2007)
17. J Mairal, F Bach, J Ponce, Task-driven dictionary learning. Pattern Anal. Mach. Intell, IEEE Trans. **34**(4), 791–804 (2012)
18. J Yang, J Wang, T Huang, Learning the sparse representation for classification, in *2011 IEEE International Conference on Multimedia and Expo (ICME)*, (Barcelona, 11-15 July 2011), pp. 1–6

doi:10.1186/1687-5281-2013-52
Cite this article as: Yu *et al.*: **Automated identification of animal species in camera trap images.** *EURASIP Journal on Image and Video Processing* 2013 **2013**:52.