EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

An automated chimpanzee identification system using face detection and recognition

Alexander Loos[1*] and Andreas Ernst[2]

Abstract

Due to the ongoing biodiversity crisis, many species including great apes like chimpanzees are on the brink of extinction. Consequently, there is an urgent need to protect the remaining populations of threatened species. To overcome the catastrophic decline of biodiversity, biologists and gamekeepers recently started to use remote cameras and recording devices for wildlife monitoring in order to estimate the size of remaining populations. However, the manual analysis of the resulting image and video material is extremely tedious, time consuming, and cost intensive. To overcome the burden of time-consuming routine work, we have recently started to develop computer vision algorithms for automated chimpanzee detection and identification of individuals. Based on the assumption that humans and great apes share similar properties of the face, we proposed to adapt and extend face detection and recognition algorithms, originally developed to recognize humans, for chimpanzee identification. In this paper we do not only summarize our earlier work in the field, we also extend our previous approaches towards a more robust system which is less prone to difficult lighting situations, various poses, and expressions as well as partial occlusion by branches, leafs, or other individuals. To overcome the limitations of our previous work, we combine holistic global features and locally extracted descriptors using a decision fusion scheme. We present an automated framework for photo identification of chimpanzees including face detection, face alignment, and face recognition. We thoroughly evaluate our proposed algorithms on two datasets of captive and free-living chimpanzee individuals which were annotated by experts. In three experiments we show that the presented framework outperforms previous approaches in the field of great ape identification and achieves promising results. Therefore, our system can be used by biologists, researchers, and gamekeepers to estimate population sizes faster and more precisely than the current frameworks. Thus, the proposed framework for chimpanzee identification has the potential to open up new venues in efficient wildlife monitoring and can help researches to develop innovative protection schemes in the future.

Keywords: Wildlife monitoring; African great apes; Face and facial feature detection; Individual identification

1 Introduction

According to the International Union for Conservation of Nature (IUCN), about 22% of the mammal species worldwide are threatened or extinct [1]. The current biodiversity crisis is observed all over the world. Primates are hit by the crisis and belong to a species that is severely endangered. Walsh et al. [2] reported a decrease of ape populations in western Equatorial Africa by more than a half between 1983 and 2000. A similar survey was done by Campbell et al. [3]. They observed a 90% decrease of chimpanzee sleeping nests in Côte d'Ivoire between 1990 and 2007.

Those agitating results demonstrate the urgent need to intensify close surveillance of this threatened species. Many protective areas have already been established. However, effectively protecting the animals requires a good knowledge of existing populations and changes of population sizes over time. Individual identification of animals is not only a prerequisite for measuring the success of implemented protection schemes but also for many other biological questions, e.g., wildlife epidemiology and social network analysis. However, it is a labor-intensive task to estimate population sizes in the wild. Therefore, noninvasive monitoring techniques that take advantage of automatic camera traps are currently under development, and the number of published studies that use camera traps or autonomous recording devices is highly

*Correspondence: alexander.loos@idmt.fraunhofer.de
[1] Audio-Visual Systems, Fraunhofer IDMT, 98693 Ilmenau, Germany
Full list of author information is available at the end of the article

increasing [4]. However, the collected data are still evaluated manually which is a time- and resource-consuming task. Consequently, there is a high demand for automated algorithms to analyze remotely gathered video recordings. Especially so-called capture-mark-recapture methods, commonly used in ecology, could benefit from an automated system for identification of great apes.

This paper shows that technology developed for human face detection and identification can provide substantial assistance in evaluating data gathered by camera traps. We summarize and extend our previous work from [5-9] on face detection and individual identification of African great apes for wildlife monitoring and present an automated framework to detect and subsequently identify free-living as well as captured chimpanzee individuals in uncontrolled environments.

Some aspects of this paper have been published in our previous work. We extended our approaches from [6] and [7] to improve the system's robustness against pose variations, difficult lighting conditions, and partial occlusions [8]. However, in this paper we present a complete system for chimpanzee photo identification including face detection, face alignment, and face recognition. We significantly improve previous approaches by fusing global and local descriptors in a decision-based manner.

While global descriptors represent the whole appearance of a chimpanzee's face, the local features around certain facial fiducial points are more robust against local changes as they only encode detailed traits of the corresponding point of interest. Furthermore, it is well known from psychophysics and neuroscience that both holistic and local information are crucial for perception and recognition of faces. Starting from the assumption that a combination of global and local descriptors should improve the performance and robustness of the system, we use a decision fusion scheme to combine their strengths. We show that global feature vectors obtained by Gabor features in combination with speeded-up robust features (SURF) [10] as local face representation achieve promising results in the new field of face recognition of great apes and clearly outperform the system presented in our previous work. For evaluation we use two realistic real-world datasets of chimpanzees, gathered in the zoo and in the field. In summary, this paper contains three main contributions:

1. Presentation of an automated framework for primate photo identification including face detection, face alignment and lighting normalization, as well as identification.
2. Extension and improvement of our previous work to achieve better performance and more robustness against pose variation, lighting conditions, facial expressions, noncooperative subjects, and even partial occlusion by branches or leafs.
3. Evaluation of the proposed system on two realistic real-world datasets of free-living and captured chimpanzee individuals gathered in uncontrolled environments.

The outcome of this paper builds the basis of an automated system for primate identification in photos and videos, which could open up new venues in efficient wildlife monitoring and biodiversity conservation management.

The remaining paper is organized as follows: In the subsequent section, we give a short recap of the existing work in the field of animal detection and identification and our own previous work. A detailed description of the proposed system, including face and facial feature detection, face alignment, and individual identification is presented in Section 3. We thoroughly evaluate our system on two datasets of free-living and captive chimpanzees in Section 4 using an open-set identification scheme. Finally, in Section 5, we conclude this paper and give further ideas of improvement.

2 Related work

The field of computer vision and pattern recognition has been an active research field for years. Even though automatic image and video processing techniques become more and more important for the detection and identification of animals, only few publications do exist dealing with that topic. In this section we give a brief overview of the existing technologies for the detection and identification of animals and briefly review face detection and recognition technologies developed for human identification.

2.1 Visual detection

Automatic face detection has been an important research area for many years now and has extensively been done for human faces. Rowley et al. [11] published good results with a neural network-based face detector more than 10 years ago. However, the system was not real-time capable at that time. Some years later, Viola and Jones [12] developed and published the probably best-known algorithm for real-time object detection. It uses AdaBoost [13] for feature selection and learning and benefits from the integral image to extract Haar-like features very fast. Numerous improvements and variants have been published in the literature afterwards [14-16].

Whereas plenty of work has already been done in the field of human face detection, only few publications can be found that deal with automatic detection, tracking, and analysis of animals in videos. Wawerla et al. [17] describe a system to monitor the behavior of grizzly bears at the arctic circle with camera traps. They use motion

shapelet features and AdaBoost to detect bears in video footage. Burghardt and Calic [18] worked on the detection and tracking of animal faces based on the Viola-Jones detector and a low-level feature tracker. They trained the system on lion faces and showed that the results can be used to classify basic locomotive actions. Spampinato et al. [19,20] proposed a system for fish detection, tracking, and species classification in natural underwater environment. They first detect fishes using a combination of a Gaussian mixture model and moving average algorithms. The detected objects are then tracked using an adaptive mean shift algorithm. Finally, species classification is performed by combining texture and shape features to a powerful descriptor.

2.2 Visual identification

One of the most established and well-studied approaches for face recognition are appearance-based methods. Here the two-dimensional gray-level images with size $w \times h$ are represented as vectors of size $n = w \cdot h$. Thus, often simple pixel-based features are used as face descriptors. Since this high-dimensional feature space is too large to perform fast and robust face recognition in practice, dimensionality reduction techniques like principal component analysis (PCA) [21], linear discriminant analysis (LDA) [22], or locality preserving projections (LPP) [23] can be used to project the vectorized face images into a smaller dimensional subspace. These methods are often referred to as Eigenfaces, Fisherfaces, and Laplacianfaces, respectively. Recently, a random projection has also been successfully used for face recognition in combination with a sparse representation classification (SRC) scheme [24]. Random projection matrices can simply be generated by sampling zero-mean independent identically distributed Gaussian entries. This approach was extended by [25]. The authors suggest to use Gabor features instead of pixel-based features, which greatly improve the recognition accuracy while at the same time reduce the computational cost when dealing with occluded face images.

While biometric identification of humans has been an active research topic for decades, individual recognition of animals has only been addressed in the recent past. Ardovini et al. [26] for instance proposed a system for semiautomatic recognition of elephants from photos based on shape comparison of the nicks characterizing the elephant's ears. A similar approach was presented by Araabi et al. [27], who proposed a string matching method as part of a computer-assisted system for dolphin identification from images of their dorsal fin. Also Burghardt et al. [28,29] presented a fully automatic system for penguin identification. After a penguin has been detected, unique individual-specific spot patterns on the penguin's coat are used for identification. More recently a method called StripeCodes for zebra identification was published by Lahiri et al. [30]. The authors claim that their algorithm efficiently extracts simple image features used for the comparison of zebra images to determine whether the animal has been observed before or not.

To the best of our knowledge, the problem of nonhuman primate identification has not yet been addressed by other researchers so far.

2.3 Own work

The aforementioned approaches use characteristic coat patterns or other individually unique biometrics like the pattern of fur and skin as well as unique nicks in ears or dorsal fins to distinguish between individuals. Unfortunately, such an approach is often infeasible for the identification of great apes since unique coat markings are not existent or cannot be used because of the limited resolution of video recordings.

Based on the assumption that humans and our closest relatives share similar properties of the face, we suggested to use and adapt face recognition techniques, originally developed to recognize humans, for the identification of great apes within the SAISBECO project (http://www.saisbeco.com). In [5] we showed that state-of-the-art face recognition techniques are capable to also identify chimpanzees and gorillas. Based on these results, we significantly improved the performance of the proposed system by using Gabor features in combination with LPP for dimensionality reduction in [6]. The SRC scheme was used to assign identities to the facial images. Although the results of [6] are very promising, the accuracy of the system drops significantly if nonfrontal face images are used for testing. Another drawback is our assumption that faces and facial feature points were already detected properly for alignment and recognition. We overcame the latter issue by combining face and facial feature detection as well as face recognition and presented an automated identification system for chimpanzees in [7]. However, we only used simple pixel information in the recognition part of the proposed system. Thus although the achieved results were very promising for a first approach, the accuracy of the system was limited due to the lack of robustness against difficult lighting situations, pose, partial occlusion, and the various number of occurring expressions.

In this paper we show how to overcome this limitation by using more sophisticated face descriptors in combination with a powerful feature space transformation technique. By combining global and local features, the system's performance and robustness against above-mentioned situations can be further increased [8]. However, this technique has never been used within a complete identification framework for great apes including face detection, face alignment, and face recognition. Therefore, in this paper we propose, design, and evaluate an automated

face detection and recognition system for chimpanzees in wildlife environments.

3 Proposed system

Figure 1 gives an overview of the proposed system. It comprises three main components. In the first step chimpanzee faces in images are found and the eyes are located within the face regions. In the second component we apply several pre-processing steps like face alignment and lighting normalization to ensure comparability of the facial images across the entire database and improve the systems robustness against lighting changes. The third and last step recognizes the detected and normalized faces and assigns identities to them. The following subsections explain those three parts in more detail.

3.1 Face and facial feature detection

Detection of primate faces and localization of the facial feature points in images are necessary for the following individual identification. Our system uses detection models with multiple classification stages of increasing complexity as shown in Figure 2. Each stage can early reject a classification window as a nonface to decrease the computational complexity. All stages comprise a feature extractor and a classifier. The feature extractors use one out of three feature types. The number of features that are used in each stage is selected empirically. The first stages of the model use few simple features that can be calculated very fast and enable real-time processing. Subsequent stages apply more complex and distinctive features. They have only minor impact on the processing speed because only few classification windows reach those complex stages. Each classifier consists of look-up tables that have been built in an off-line training procedure using Real-AdaBoost [31]. Our system uses three types of features that are illumination invariant and robust against various lighting conditions. All features are solely based on gray value images and thus enable the system to process infrared images as well.

The first feature type describes the local gradient direction. Sobel kernels of size 3×3 extract the gradient s_x and s_y in x- and y-direction, similar to [32]. In homogeneous regions where s_x and s_y equals 0, the final feature is encoded as 0; otherwise, the feature encodes the result of $\text{atan2}(s_y, s_x)$ quantized to the range $1 \ldots q$. Experiments indicated that 35 is a good choice for q and results in a quantization interval of slightly more than 10°. We use *census* features [33] (also known as local binary patterns [34]) as a second feature type. These features describe the local brightness changes within a 3×3 neighborhood. The center pixel intensity is compared with its eight neighbors. The result is encoded as an 8-bit string that shows which neighboring pixels are less bright than the center pixel. The 3×3 local features are complemented by the third feature type that includes enlarged areas. Therefore, we encode structures by resized versions of census features that are calculated on image regions of $3u \times 3v$ pixels. These *structure* features are a superset of the census features. Nevertheless, considering census features separately is justified well in terms of processing speed because they can be calculated much faster for the whole image.

The distinction between pixel-based gradient features, census features, and region-based structure features is important for real-time requirements. Pixel-based features are calculated beforehand for the whole image and reused when sliding the analysis window over the image. Region-based features have to be calculated separately for each analysis window. Pixel-based features are suited for fast candidate search, whereas more significant region-based features improve the performance of candidate verification. We choose a model size of 24×24 pixels that is commonly used for human face detection and obtain 484 gradient features, 484 census features, and 8,464 structure features. The first stages offer a quick candidate search and the final stages provide a more accurate but slower classification. The training procedure starts with randomly chosen nonface data for the initial stage. Following stages are trained with

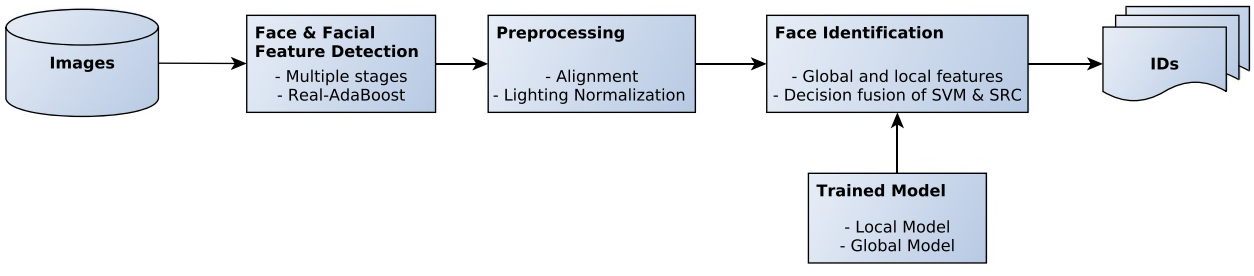

Figure 1 Overview of the proposed system. The figure depicts the overview of the proposed system for chimpanzee identification. After all possible faces in an image are detected, each face is aligned using a projective transform to make the faces comparable across the entire dataset. A histogram equalization is also applied in this step to improve the system's robustness against changing lighting conditions. In the final stage of our framework, the detected and aligned faces are identified using a combination of global and local features.

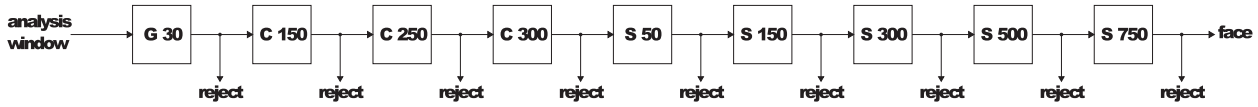

Figure 2 Setup of the face detection model. The face detection model comprises nine stages with different feature types and feature counts. The letters abbreviate 'G' for gradient, 'C' for census and 'S' for structure features. The first stages apply fast pixel-based features for the candidate search; the last stages use structure features for verification.

nonface data that are gathered by bootstrapping the model on images without ape faces. More details about the training procedure can be found in our previous work [9].

A 3×3 mean filter reduces noise in the input image. We resize the filtered image with different scaling factors and generate an image pyramid to detect faces of arbitrary size. The detection model of size 24×24 pixels analyzes each pyramid level with a coarse to fine search to further improve speed. Therefore, the detection model is shifted with a step size of about 6 pixel across each pyramid level. The neighborhood of a grid point is scanned more thoroughly only if the grid point produced a high face correlation score.

After the face detection process, we apply a subsequent eye search in all detected face regions with the same algorithms. We trained a detection model for each eye with a reduced size of 16×16 pixels. Only the eye regions were cut out from the annotated training data for this purpose. The eye models are simpler and less powerful compared to the face model and comprise five stages only and less features, because searching within face regions will lead to few false positives. Selected areas in all face regions around the left and right eye are scanned with the appropriate eye model in different scaling levels. Fixed eye markers of the face model are used if an eye could not be detected by the eye search.

3.2 Face alignment and lighting normalization

A very crucial step to achieve good performance in the subsequent face recognition task is the alignment of the detected faces. Based on the automatically detected eye coordinates, we first rotate the facial image into an upright position such that the eyes lie on a horizontal line. If coordinates for the center of the mouth are not available, we estimate the locations of the left and right corner of the mouth based on the eye coordinates only. However, if the location of the center of the mouth is provided, we can calculate the position of the mouth's corners more precisely, which will lead to a better alignment. Based on the locations of the left and right eye as well as the left and right corners of the mouth, we are then able to apply a projective transform to finally align the ape's face. This step ensures that facial features like eyes, nose, and mouth are located nearly at the same coordinates throughout the entire dataset. Consequently, this guarantees that extracted visual facial features are comparable for all faces. Figure 3 illustrates the applied face alignment procedure for an example image. After converting the aligned face image into gray scale, we apply a simple histogram equalization for lighting normalization and contrast adjustment.

3.3 Individual identification

The individual identification is the main part of the proposed system and consists of three steps: feature extraction, feature space transformation, and classification. In the first step we extract global as well as local visual features that are both well suited for discrimination. As those descriptors are too high dimensional to perform fast and robust face recognition in practice, we apply a feature space transformation technique called LPP [23] to achieve a lower dimensional subspace with only little loss of information that is important for identification. These lower dimensional feature vectors are then used for classification. After classifying the global and local feature vectors separately, we apply a decision fusion technique to get the final result.

3.3.1 Feature extraction

Since global features gather holistic information of the face and local descriptors around facial points represent intrinsic factors, both should be used for classification. Additionally, it has been reported in the literature that different representations misclassify different patterns [35]. Therefore, various features offer complementary information which can be used to improve the recognition results. As for global features we propose to use Gabor features, which are known to perform well in pattern recognition tasks. The complimentary local descriptor is SURF, a powerful visual descriptor of interest points in an image.

Gabor descriptor Gabor features are known to perform well in face recognition and pattern recognition tasks in general [36-38]. They are extracted by convolving the gray-level input image $I(z)$ with a set of Gabor kernels $\psi_{\mu,\nu}(z)$ as

$$G_{\mu,\nu}(z) = I(z) * \psi_{\mu,\nu}(z), \tag{1}$$

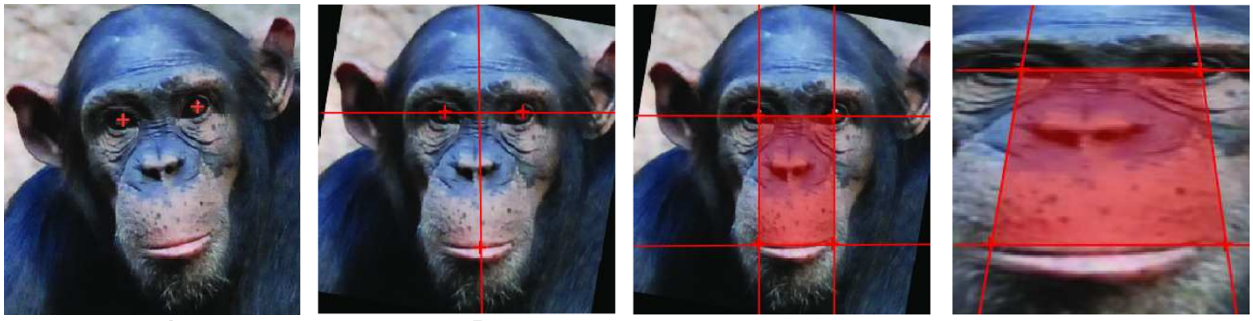

Figure 3 Face alignment. The face alignment procedure for an example image. Based on the detected eye coordinates the position of the mouth is estimated **(A)**. After rotating the facial image into an upright position **(B)**, such that both eyes lie on a horizontal line, the left and the right corner of the mouth is estimated **(C)**. Based on these four points a projective transformation is applied **(D)**. This ensures that facial features like eyes, nose, and mouth are located approximately at the same positions throughout the entire dataset, which is a prerequisite for accurate identification.

where $G_{\mu,\nu}(z)$ is the output image at orientation μ and scale ν at pixel location $z = (x, y)$. Complex Gabor kernels are defined as

$$\psi_{\mu,\nu}(z) = \frac{\|k_{\mu,\nu}\|^2}{\sigma^2} e^{\frac{-\|k_{\mu,\nu}\|^2 \|z\|^2}{2\sigma^2}} \left[e^{ik_{\mu,\nu} z} - e^{\frac{-\sigma^2}{2}} \right], \tag{2}$$

where the wave vector $k_{\mu,\nu}$ is defined as $k_{\mu,\nu} = k_\nu e^{i\theta_\mu}$ with $k_\nu = \frac{k_{\max}}{f^\nu}$ and $\theta_\mu = \frac{\pi\mu}{8}$. The maximum frequency is denoted as $k_{\max}$ and f is the spacing between kernels in the frequency domain. Furthermore, σ represents the ratio of the Gaussian window to the wavelength.

In general, $G_{\mu,\nu}(z)$ is complex and can be rewritten as $G_{\mu,\nu}(z) = M_{\mu,\nu}(z) e^{i\theta_{\mu,\nu}(z)}$, where $M_{\mu,\nu}(z)$ denotes the magnitude, and $\theta_{\mu,\nu}(z)$ the phase at pixel location z. Since the magnitude contains the local energy variation in the facial image, $M_{\mu,\nu}$ is used as feature, while $\theta_{\mu,\nu}(z)$ is ignored for further processing. Finally, the overall feature vector is constructed as

$$x_{\text{GABOR}} = \left(m_{0,0}^{(\rho)}, m_{0,1}^{(\rho)}, \cdots, m_{1,0}^{(\rho)}, \cdots, m_{K,L}^{(\rho)} \right), \tag{3}$$

where $m_{\mu,\nu}^{(\rho)}$ is a column vector representing the normalized and vectorized version of the magnitude matrix $M_{\mu,\nu}$ which was down-sampled by factor ρ.

For feature extraction we use five scales and eight orientations for the generation of Gabor kernels with size of 31×31. We chose to set $k_{\max} = \frac{\pi}{2}, f = \sqrt{2}$, and $\sigma = \pi$. After convolving an image with the resulting 40 Gabor wavelets, we down-sample the magnitude matrix $M_{\mu,\nu}$ by a factor of $\rho = 8$ by using a bilinear interpolation.

SURF descriptor SURF is a fast and robust scale- and rotation-invariant interest point detector and descriptor. It was first published by Bay et al. in 2008 [10]. In this task we already know the position of the interest points so that we only refer to the descriptor part of SURF in this paper. In the following we briefly describe the main ideas of SURF. A more detailed description including the detection of interest points can be found in [10]. As claimed by the authors, the standard version of SURF is several times faster, more compact, and, at the same time, more robust against certain image transformations than comparable local descriptors like scale invariant feature transform (SIFT) [39]. Similar to SIFT and its variants, SURF describes the distribution of intensity content within a certain neighborhood around the interest point. However, instead of using gradient information directly, SURF uses first-order Haar wavelet responses in x and y-direction. For efficiency, SURF exploits integral images which drastically reduces processing time while at the same time improves the robustness of the resulting descriptor. In order to increase the robustness against rotation, usually the first step of feature extraction is to identify a reproducible orientation for the interest point. The dominant orientation can be found by calculating the sum of the Gaussian-weighted Haar wavelet responses using a sliding window around a circular region around the interest point. The next step is to construct a square region with correct orientation symmetrically around the interest point. This region is then split into 4×4 subregions. Finally, the feature vector can be calculated by again using Haar wavelet responses weighted with a Gaussian kernel which is centered at the particular interest point. The horizontal and vertical wavelet responses, dx and dy, as well as their absolute values are summed up over each subregion to construct the final feature vector of size 64

$$x_{\text{SURF}} = \left(\sum dx, \sum dy, \sum |dx|, \sum |dy| \right). \tag{4}$$

We extract SURF descriptors on six facial fiducial points which are calculated based on the detected eye markings. Figure 4 shows the location of the facial markings. Note that for the local feature extraction, we just rotate

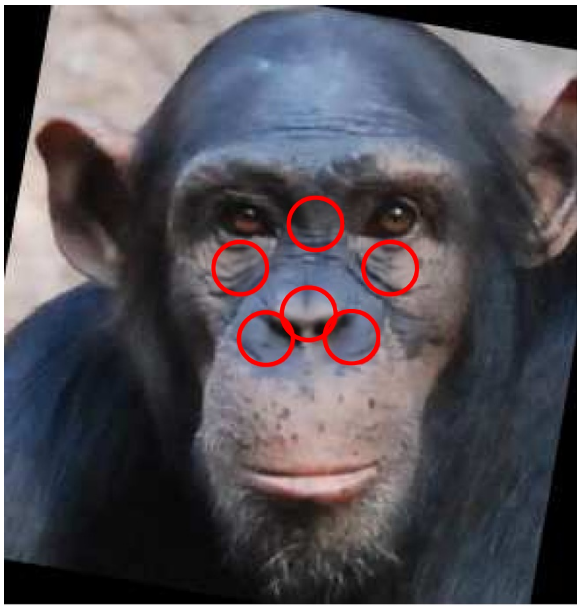

Figure 4 Local feature extraction. The positions of the applied SURF descriptor for local feature extraction. Three out of six positions for local feature extraction are located under and between both eyes; the remaining three interest points are situated on the nose tip and both nostrils. The mouth region is not used for feature extraction because we especially noticed that this region is often subject to occlusion and facial expressions.

the faces into an upright position for alignment instead of applying a projective transform to prevent unnatural distortion of local regions and then resize the facial image to 64×64 pixels. Since we already performed this step during face alignment as discussed in Section 3.2, we do not use the rotation-invariant version of the SURF descriptor. This saves computation time because it is not necessary anymore to identify the main orientation of the interest point beforehand. As stated in [10], the upright version of SURF is faster to compute and can increase distinctiveness while maintaining a certain robustness against small rotation angles of up to $\pm 15°$. Furthermore, we only need to compute the descriptor in one particular scale because we previously resized every face image to a fixed size. This also makes the feature extraction step more efficient and facilitates real-time performance of the final system. Based on the assumption that wrinkle patterns under and between the eyes are unique across individuals and useful for identification, the first three points are located under the left and right eye, as well as between both eyes. Furthermore, we assume that the area around the nose is well suited for discrimination. Therefore, we use the tip of the nose as well as the left and the right nostril as additional locations for local feature extraction. We do not extract information out of the mouth region because we noticed that this area is often subject to occlusion and deformation because of eating and facial expressions. Extracting features out of this region would lead to a high intra-class variance and would hamper classification.

3.3.2 Feature space transformation

The goal of many feature space transformation techniques is to project the N high-dimensional feature vectors $\{x_1, \cdots, x_N\}$ of size n into a smaller dimensional subspace of size m using a unitary projection matrix $W \in \mathbb{R}^{n \times m}$

$$y_k = W^T x_k; \quad \text{with} \quad x_k \in \mathbb{R}^{n \times 1}, y_k \in \mathbb{R}^{m \times 1}, m \leq n. \tag{5}$$

The resulting feature vectors $y_k \in \mathbb{R}^{m \times 1}$, with $k = 1, \cdots, N$, can then be used for classification.

LPP [23] assumes that the feature vectors reside on a nonlinear submanifold hidden in the original feature space. LPP tries to find an embedding that preserves local information by modeling the manifold structure of the feature space using a nearest-neighbor graph. First, an adjacency graph G is defined. An edge is put between two nodes k and j if they belong to the same class C.

LPP will try to optimally preserve this graph when choosing projections. After constructing the graph, weights have to be assigned to the edges. Therefore, a sparse symmetric matrix S of size $N \times N$ is created with $S_{k,j}$ having the weight of the edge joining vertices k and j, and 0 otherwise. The weights are calculated as follows:

$$S_{k,j} = \begin{cases} e^{-\frac{\|x_k - x_j\|^2}{2*\sigma^2}}, & \text{if} \quad C(x_k) = C(x_j) \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Here, σ denotes a constant for normalization which we set to 100 in our system. The objective function of LPP is defined as

$$w_{\text{opt}} = \min \sum_{kj} (y_k - y_j)^2 S_{k,j}. \tag{7}$$

Following some simple algebraic steps, it is possible to show that Equation 7 finally results in a generalized eigenvalue problem:

$$XLX^T w = \lambda XDX^T w, \tag{8}$$

where D is a diagonal matrix whose entries are column sums of S and $L = D - S$ is the so-called Laplacian matrix. The kth column of matrix X is x_k.

The projection matrix W is constructed by concatenating the solution to the above equation, i.e., the column vectors of $W_{\text{LPP}} = [w_1, \cdots, w_m]$ are ordered ascendingly according to their eigenvalues. Usually, the original features are first projected into the PCA subspace before

applying LPP by deleting the smallest principle components. Thus, the final embedding is as follows:

$$W_{\text{final}} = W_{\text{PCA}} W_{\text{LPP}}. \tag{9}$$

Details about the algorithm and the underlying theory can be found in [23].

3.3.3 Classification

Sparse representation classification For the classification of global features, we use the SRC paradigm developed by Wright et al., which is known to perform well for face recognition [24,25].

Let $\tilde{A} \in \mathbb{R}^{m \times l}$ be the normalized matrix of training samples transformed into the feature space and $\tilde{t} \in \mathbb{R}^{m \times 1}$ be the normalized transformed feature vector of the test image, where m is the dimensionality of the feature space and l the number of training samples. Classification can be done by first solving a convex optimization problem via l_1-norm minimization:

$$\hat{p} = \arg\min_{p} \|p\|_1 \quad \text{subject to} \quad \tilde{t} = \tilde{A}p, \tag{10}$$

where $p \in \mathbb{R}^{l \times 1}$ is a sparse coefficient vector whose entries only associated with the ith class should be 1 and the rest be 0. In other words, we try to represent the feature vector $\tilde{t}$ of the test image as a linear combination of the training samples of the same class. Therefore, $\tilde{t}$ is assigned to the object class that minimizes the residual $r_i(\tilde{t})$ between $\tilde{t}$ and $\tilde{A}(\delta_i \odot \hat{p})$ such that

$$\min_i r_i(\tilde{t}) = \|\tilde{t} - \tilde{A}(\delta_i \odot \hat{p})\|_2, \tag{11}$$

where $\odot$ denotes the elementwise multiplication known as Hadamard product. The vector $\delta_i \in \mathbb{R}^{l \times 1}$ is called the characteristic function of class i. It is a filter vector which is 1 for all training samples of class i and 0 elsewhere. A detailed description of SRC can be found in [24].

Support vector machines In the proposed system, we use a support vector machine (SVM) [40] for the classification of local features. SVM is a discriminative classifier, attempting to generate an optimal decision plane between feature vectors of the training classes. Often, the classification with linear separation planes is not possible in the original feature space for real-world applications. Using a

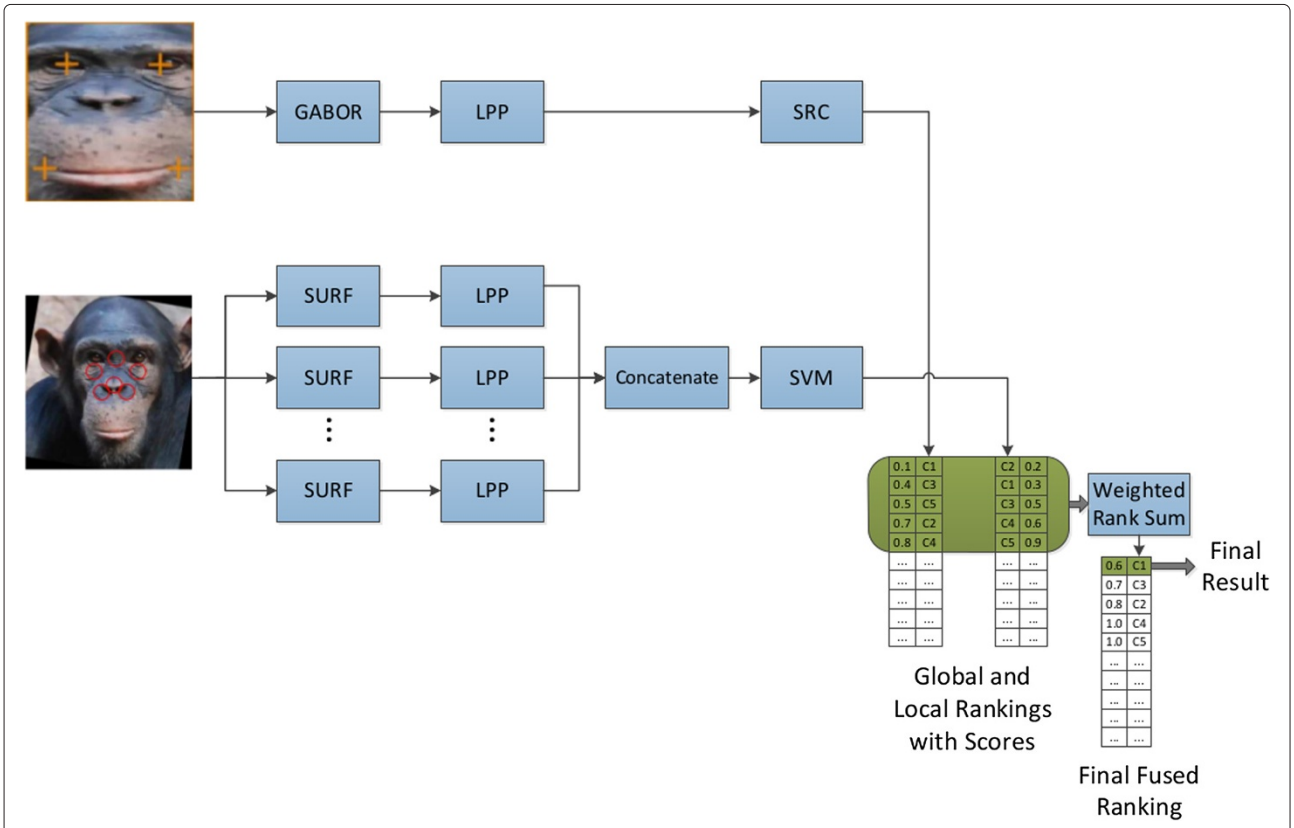

Figure 5 Proposed parallel fusion scheme. This figure shows the parallel fusion scheme to combine the results of global Gabor features and local SURF descriptors. Both global and local features are first projected into a smaller dimensional subspace using LPP. Note that we transform each SURF feature separately into the feature space before concatenating the resulting vectors to a comprehensive local feature vector. The global feature is classified using SRC while the local feature vector is classified by SVM with RBF kernel. The ranked results are then combined using the decision fusion rank sum method explained in Section 3.3.4 to obtain the final result.

Table 1 Overview of the datasets we used in our experiments

Dataset	Images	Individuals	Faces	∑ Pixels (MP)
ChimpZoo	2,617	24	598	6,403
ChimpTaï	3,905	71	1,432	5,409

so-called kernel trick, the feature vectors are transformed into a higher dimensional space in which they can be linearly separated. We use a radial basis function (RBF) as kernel in this paper.

3.3.4 Decision fusion

The decision fusion paradigm we use in this paper was influenced by ideas of [41]. A parallel ensemble classifier which fuses the rank outputs of different classifiers is used to combine the results of local and global features. In contrast to the parallel fusion scheme proposed in [41], where only a single weighting function $w(\mathcal{R}) = \mathcal{R}^c$ for rank $\mathcal{R}$ and constant c is used as nonlinear rank sum method, we weight the results of both classifiers using different weighting functions for every classifier. Additionally, the confidences of each classifier can be taken into account when generating the weighting function $w(\mathcal{R}) = e^{s(\mathcal{R})}$, where $s(\mathcal{R})$ represents the confidence of SRC or SVM for rank $\mathcal{R}$. For SRC we use the vector of residuals from Equation 11 as confidence measure, while for SVM the probability estimates of LibSVM [40] can be utilized. The probability estimates can simply be converted into match scores by negating the probabilities. Details on the estimation of probabilities for SVM can be found in [42]. The final score vector $s_f \in \mathbb{R}^{C\times 1}$, where C is the number of classes, is then simply the sum of both weighting functions: $s_f = w_{\mathrm{SRC}} + w_{\mathrm{SVM}}$. Finally, s_f is ordered ascendingly to obtain the final result.

Figure 5 illustrates the parallel fusion scheme we use in this paper. Note that for every of the six facial interest points, we transform the resulting SURF descriptors separately into a smaller dimensional subspace before concatenating them to get the final local feature vector.

4 Experiments and results

4.1 Dataset description

Due to the lack of publicly available face databases of chimpanzees, we use self-assembled annotated datasets of captive as well as free-living individuals from the Zoo Leipzig, Germany, and the Taï National Park, Côte d'Ivoire, Africa, respectively. For benchmark purposes, the license rights for both datasets can be purchased over our project website http://www.saisbeco.com. Both datasets were annotated by experts. Table 1 gives details about both datasets that were used in our experiments.

Example images for both datasets with detected faces can be seen in Figure 6. To have a valid ground truth for evaluation, the position of the head, eyes, and mouth was annotated. Metadata were also assigned to every annotated face such as gender, age and the name of the individual. The experts used our proprietary annotation tool for this purpose. This tool allows the annotation of face regions in the image along with related facial marker points. Moreover, metadata can be assigned to all faces by additional attributes. The annotations are stored separately for each image in a XML file. More details about the annotation tool can be found in our previous work [43]. Figure 7 shows detected faces of one individual for the ChimZoo dataset (Figure 7A) and the ChimpTaï dataset (Figure 7B), respectively. It is obvious that both datasets are very challenging for the recognition task because detected faces of one single individual can have a variety of poses, expressions, lighting situations and even partial occlusion by branches or leafs. Thus, the algorithm used for identification is required to be robust against that kind of variations to achieve sufficient recognition results.

4.2 Evaluation measures and experiment design

Since the face detection stage will produce false-positive detections, we decided to use an open-set identification

Figure 6 Examples of detected faces. Example images of detected faces for the ChimpZoo dataset **(A)** and the ChimpTaï dataset **(B)**. The region of the successfully detected faces and eyes are marked (in green lines). Additionally, the species is automatically assigned to the face.

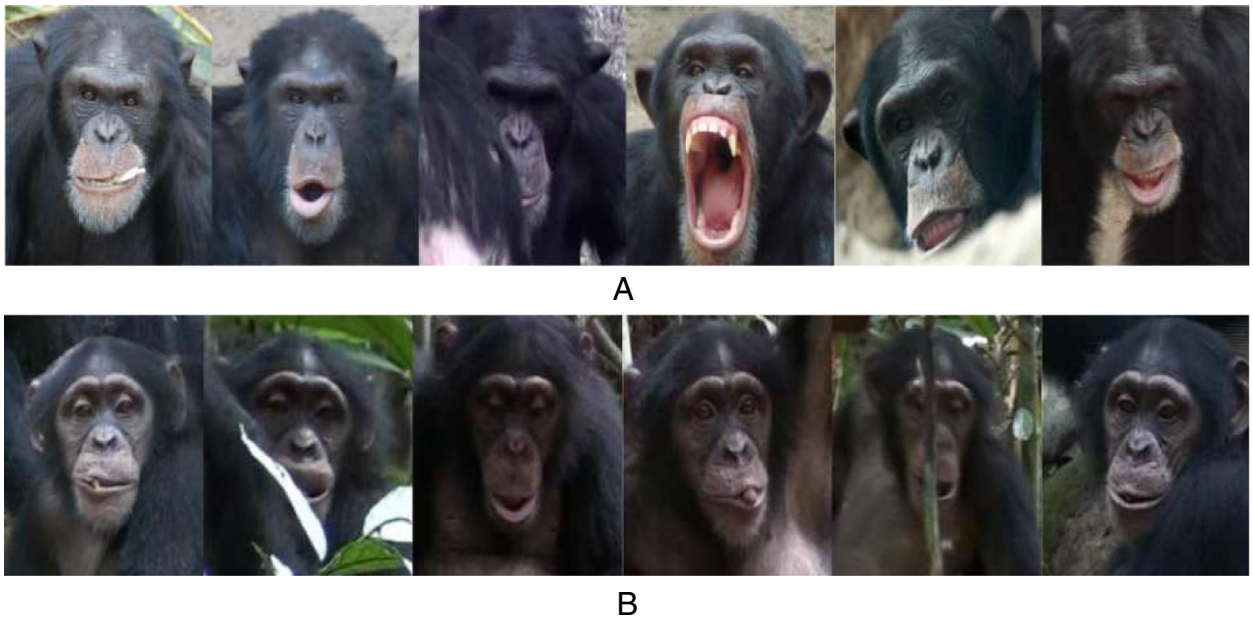

Figure 7 Detected faces of one individual per dataset. Detected faces of one individual of the ChimpZoo dataset **(A)** and the ChimpTaï dataset **(B)**. Both datasets are very challenging due to difficult lighting situations, facial expressions, poses, and even partial occlusion by branches or leafs.

scheme to deal with that issue. We use the performance statistics described in [44-46] to evaluate our system. In open-set identification, first the system has to decide if the probe p_j represents a sample of an individual in the gallery $\mathcal{G}$ or not. If the system decided that the individual in the probe is known to the system, then it also has to report the identity of the individual. While for a closed-set identification, the question is how many test images are correctly classified as a certain individual, two more types of errors can occur for an open-set classification. Additional to false classifications, it is also possible that the system rejects known individuals or accepts impostors. Let $\mathcal{P}_\mathcal{G}$ be the probe set that contains face images of chimpanzees in the gallery and $\mathcal{P}_\mathcal{N}$ the probe set that contains samples of chimpanzees that are not known to the system. When a probe p_j is presented to the system, a score vector $s \in \mathbb{R}^{C\times 1}$ can be calculated, where C is the number of known individuals in the database. The entries of this vector are scaled between 0 and 1. The smaller the value, the higher the confidence of the classifier. For SRC we use the vector of residuals r from Equation 11 as confidence measures, while for the proposed decision fusion technique, the combined weightings s_f can be used as score values for each class. For classification by SVM, the probabilities of the classifier can be negated to assign them to the score vector s_f.

A probe p_j is detected and identified if the minimal score $s_{\min,j}$ is below the operating threshold τ and identified correctly with $\text{rank}(p_j) = 1$. Therefore, the detection and identification rate P_{DI} at threshold τ can be calculated as

$$P_{\text{DI}}(\tau, 1) = \frac{\left|\{p_j : p_j \in \mathcal{P}_\mathcal{G}, \text{rank}(p_j) = 1, \text{and } s_{\min,j} \leq \tau\}\right|}{|\mathcal{P}_\mathcal{G}|}. \tag{12}$$

The second performance measure is the false alarm rate P_{FA}. A false alarm occurs when the minimal match score

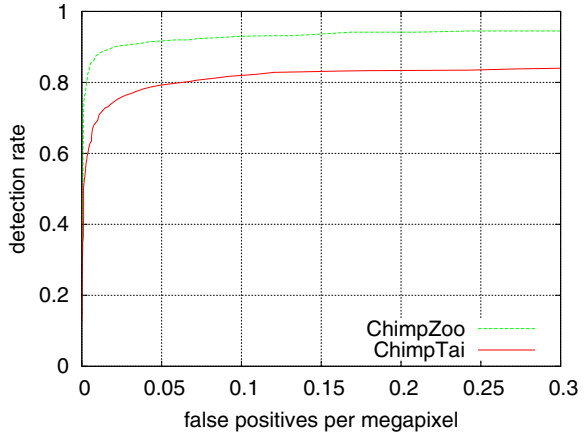

Figure 8 Face detection performance. This figure shows the ROC curve of the face detection model evaluated on the ChimpZoo and ChimpTaï dataset. The number of false detections is normalized to the total pixel sum of both datasets.

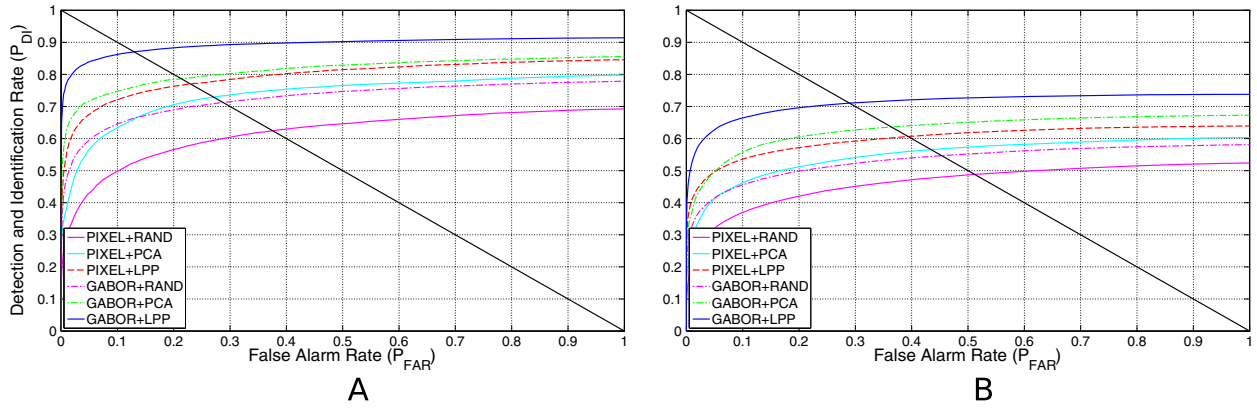

Figure 9 Results of experiment 1. ROC curves for the first experiment we conducted in this paper for the ChimpZoo dataset **(A)** and the ChimpTaï dataset **(B)**. The black solid line denotes the line of equal error. We compared globally extracted Gabor features (GABOR) with pixel-based features (PIXEL). We combined the features with three different methods for feature space transformation, random projection (RAND), Principal component analysis (PCA), and locality preserving projections (LPP). For all combinations we used the SRC algorithm for classification. It can be seen that Gabor features perform best in most of the cases and are therefore better suited for describing chimpanzee faces than simple pixel-based features. Our proposed approach (GABOR + LPP), which is denoted by the solid blue line, outperforms all the other algorithms with an equal error rate (EER) of 0.1290 and 0.2938 for the ChimpZoo and ChimpTaï dataset, respectively.

of an impostor is below the operating threshold τ. Consequently, the false alarm rate is the fraction of probes in $\mathcal{P}_{\mathcal{N}}$ that are detected as genuine individuals and is calculated as

$$P_{\mathrm{FA}}(\tau) = \frac{\left|\{p_j : p_j \in \mathcal{P}_{\mathcal{N}}, s_{\min,j} \leq \tau\}\right|}{|\mathcal{P}_{\mathcal{N}}|}. \tag{13}$$

An ideal system would have a detection and identification rate of 1.0 and a false alarm rate of 0.0, which means that all individuals are detected and classified correctly and there are no false alarms. In practice however, both measures have to be traded-off against each other. This trade-off is shown in a receiver-operating characteristic (ROC) by iteratively changing the operating threshold τ. Another important performance statistic is the equal error rate (EER). It is reached when the false alarm rate is equal to the false detection and identification rate $P_{\mathrm{FA}} = 1 - P_{\mathrm{DI}}$.

In addition to false-positive detections, one individual at a time is removed from the training set and presented it as an impostor to test the system's capability to reject unknown chimpanzees. This procedure is repeated C times, where C is the number of individuals in the dataset, such that every chimp takes the role of an impostor once. To get valid results, we additionally apply a tenfold stratified cross validation. Images of false-positive detections as well as all pictures of the unknown individual remain in the test set for all ten folds and are not used for training. We only consider detections with a minimum size of 64×64 pixels for identification, which dramatically decreases the number of false-positive detections. Furthermore, we only focus on individuals with at least five detected face images in the database to get an appropriate number of training images for each class. This limitation results in 24 individuals for the ChimpZoo and 48 subjects for ChimpTaï dataset. After aligning the detected face images as described in Section 3.2, we apply a histogram equalization for lighting normalization. To make the results comparable, we chose to have a feature dimension of 160 for all applied feature space transformation techniques. For the local SURF features, we transform the resulting feature vectors separately into a smaller dimensional subspace of size 50 for every of the six used facial fiducial points before concatenating them to the final feature vector. This results in a local feature vector of size $6 \times 50 = 300$.

4.3 System evaluation

4.3.1 Face detection

Because of the lack of publicly available face databases for chimpanzees, we trained the detection model with frontal

Table 2 EER for Gabor and pixel-based features for feature space transformation

	ChimpZoo		ChimpTaï	
EER	**Pixel**	**Gabor**	**Pixel**	**Gabor**
RAND	0.3804	0.2885	0.5171	0.4504
PCA	0.2671	0.2161	0.4363	0.3647
LPP	0.2313	**0.1290**	0.3950	**0.2938**

Equal error rates (EER) for Gabor and pixel-based features in combination with random projection (RAND), principal component analysis (PCA), and locality preserving projections (LPP) for feature space transformation. Our proposed approach is in boldface and performs best on both datasets.

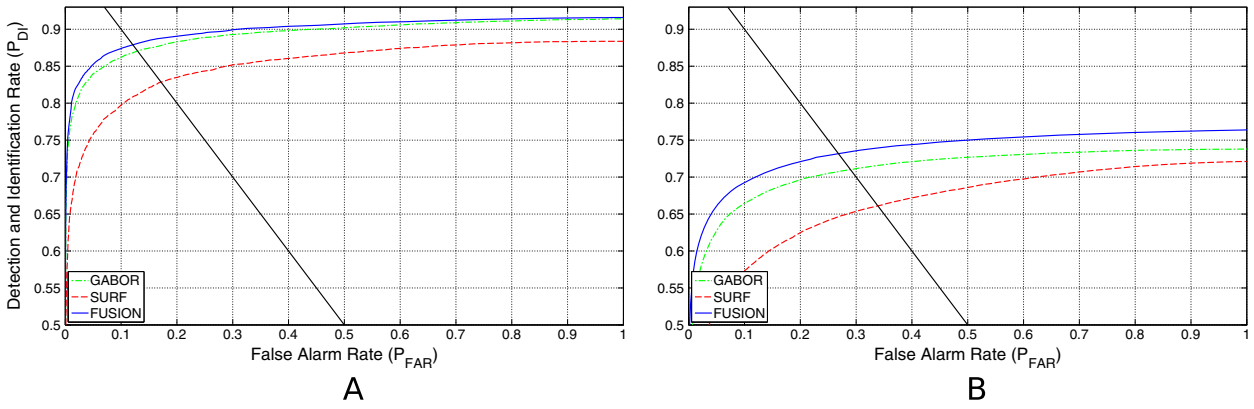

Figure 10 Results of experiment 2. Achieved results for the second experiment we conducted for the ChimpZoo dataset **(A)** and the ChimpTaï dataset **(B)**. Again, the black solid line denotes the line of equal error. The green dash-dotted line shows the result for globally extracted Gabor features in combination with LPP and SRC. The red dashed line denotes the results for the local SURF descriptors extracted around six facial feature points. Each feature vector was transformed separately into a smaller dimensional subspace using LPP before classifying the concatenated final feature vector using a SVM with an RBF kernel. Although the locally extracted features perform worse than the global descriptor, they carry important information for classification. This can be seen from the results of the decision fusion scheme (blue solid line) that outperforms algorithms based on global and local features alone with an EER of 0.1189 and 0.2669 for the ChimpZoo and the ChimpTaï dataset, respectively.

faces of the ChimpZoo dataset and evaluated it on both datasets. Figure 8 shows the results with ROC curves. The detection rate on the ChimpZoo dataset is considerably higher. This can be expected because this dataset was used for training. Moreover, it shows a higher image quality than the ChimpTaï dataset in terms of resolution and extrinsic factors like lighting conditions, contrast, and occlusion. A threshold defines the working point of the detector on the ROC curve. If we accept 0.1 false positives per megapixel in practice, the detector finds 93% and 82% of the faces in the ChimpZoo and ChimpTaï dataset, respectively.

4.3.2 Face identification

Experiment 1: influence of visual features and feature space transformation In the first experiment we want to address the question if Gabor features (GABOR) or pixel based features (PIXEL), used in our previous work [5,7], are better suited for face recognition of great apes. Furthermore, we evaluate and compare three different feature space transformation techniques: random projection (RAND) [24], PCA [21], and LPP [23]. To make the results comparable, we set the number of features to 160 for every feature space transformation method. For the face alignment procedure, we used the manually annotated facial marker points in this experiment. Figure 9 shows the results for the ChimpZoo dataset (Figure 9A) and ChimpTaï dataset (Figure 9B), respectively. The black diagonal line denotes the line of equal error. For all combinations we used the SRC algorithm for classification. As can be seen, our approach to use Gabor features as global descriptors and LPP for feature space transformation outperforms all the other approaches on both datasets (blue solid line). The EER for every algorithm and both datasets can be seen in Table 2.

Since global Gabor features in conjunction with LPP achieves the best results in the first experiment, this combination should be used for holistic face recognition for primates. However, in the next experiment we will show that this algorithm can still be enhanced by additionally using locally extracted SURF features and our proposed decision-based fusion scheme.

Experiment 2: combination of global and local features In the second experiment, we show that our proposed approach for combination of global and local features improves the performance and outperforms systems based on global or local features alone. As specified in Section 3, we use Gabor features as global face representation and SURF as local descriptors. Both features are transformed into a lower dimensional subspace using LPP. Note that we apply LPP on SURF features separately for every interest point before generating the final local feature vector. Again, for the global Gabor feature vector, we set the number of dimensions to 160 while each SURF descriptor is transformed into a feature space of size 50, resulting in a combined local descriptor of

Table 3 EER for global Gabor and local SURF features and proposed parallel decision fusion scheme

EER	Gabor	SURF	Fusion
ChimZoo	0.1290	0.1903	**0.1189**
ChimpTaï	0.2938	0.3171	**0.2669**

The proposed decision fusion method is printed in bold and performs best on both datasets.

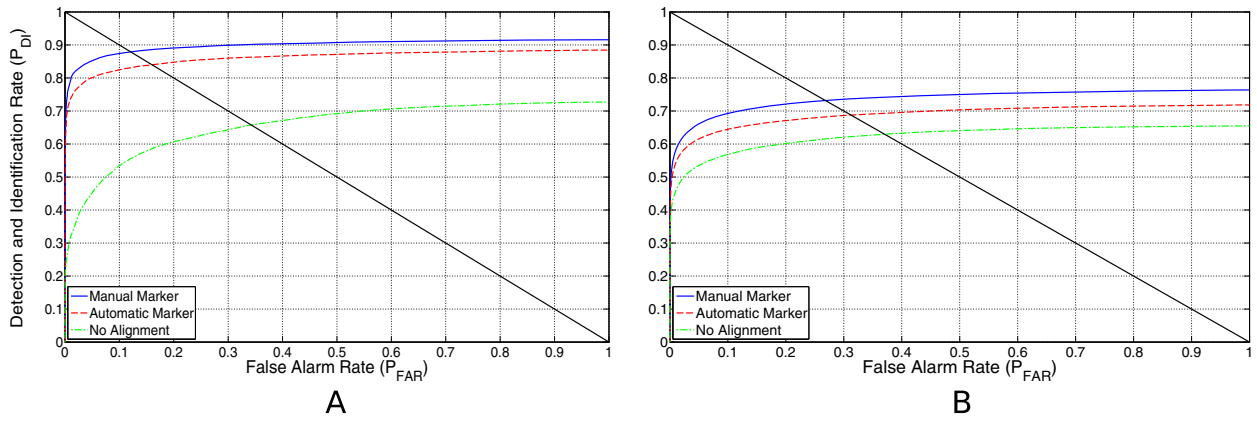

Figure 11 Results of experiment 3. ROC curves of the proposed identification system for alignment using manually annotated coordinates of both eyes and mouth (blue solid line), automatically detected eyes (red dashed line), and without applied alignment (green dash-dotted line). **(A)** Curves for the ChimpZoo dataset. **(B)** Results for the ChimpTaï dataset. The black solid line denotes the line of equal error. As can be seen, the performance of the algorithm if manually annotated facial markings were used for alignment is only slightly better than alignment using automatically detected facial feature points. The accuracy of the system drops significantly if no alignment was applied at all.

size 300. The classification is done separately for global and local features using SRC and SVM, respectively. The results for global and local features are combined in the decision-based manner we described in Section 3.3.4. Like in experiment 1, the manually annotated eye coordinates were used for alignment. Figure 10 shows the resulting ROC curves for the ChimpZoo and ChimpTaï dataset, respectively. For both datasets the global Gabor features (green dash-dotted line) perform significantly better than the local SURF descriptors (red dashed line). Nevertheless, obviously SURF descriptors encode important information to discriminate between individuals which can be seen from the results of the proposed fusion paradigm (blue solid line). The decision fusion of global and local features performs better than global and local features alone, especially for the free-living individuals (Figure 10B). The associated equal error rates can be seen in Table 3.

It is obvious that our proposed fusion scheme performs better than global and local features alone. Therefore, the idea of using the confidences of both classifiers improves the performance of the face recognition algorithm for chimpanzee faces in real-world environments.

However, we still used manually annotated eye coordinates for alignment and estimation of facial fiducial points for local feature extraction. In the final experiment we use the automatically detected facial markings for this purpose.

Experiment 3: manually annotated vs. automatically detected facial markings In the third and last experiment, we show that automatically detected facial feature points for face alignment perform almost as good as manually annotated ones. For facial feature detection, we use the algorithm described in Section 3.1. For face identification we use the proposed system that combines global and local features for recognition which performed best in experiment 2. All parameters were set as described in the previous experiment. We compare the recognition results of the system for face alignment with manually annotated feature points, automatically detected facial feature points, and if no alignment was applied at all. Figure 11 shows the ROC curves of the proposed identification algorithm for alignment using the manually annotated facial feature points (blue solid line), automatically detected markings (red dashed line), and without applied alignment (green dash-dotted line) for the ChimpZoo dataset in Figure 11A and the ChimpTaï dataset in Figure 11B. The according equal error rates can be seen in Table 4.

If manual markings were used for alignment and estimation of the facial fiducial points for local feature extraction, the proposed algorithm performs best. However, if we use automatically detected eye coordinates, the performance of the algorithms is only slightly worse than for manually annotated markings. This is because the automatic detection of eye coordinates is not always as accurate as manually detected ones. Another explanation is that for the automatically detected markings, it was only possible to estimate the coordinates for local feature extraction

Table 4 EER of the proposed identification algorithm if alignment was applied

EER	Manual markings	Automatic markings	No alignment
ChimZoo	0.1189	0.1590	0.3447
ChimpTaï	0.2669	0.3103	0.3668

EER of the proposed identification algorithm if alignment was applied using manually annotated markings, automatically detect markings, or if no alignment was applied as a pre-processing step. The EER for manually annotated markings and automatically detected feature points are very close while the performance of the algorithm drops significantly if no alignment was performed at all.

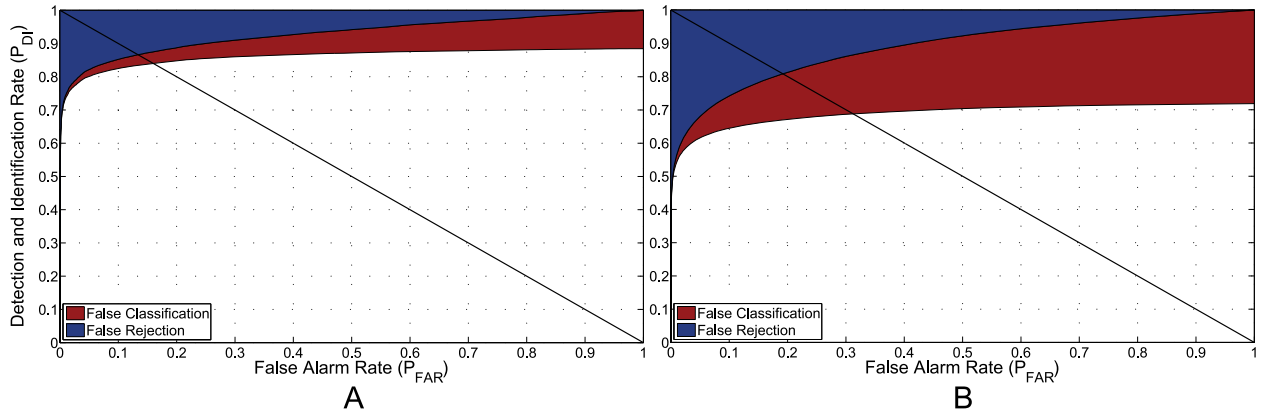

Figure 12 Contribution of P_{FR} and P_{FC} to the overall accuracy of the system. This figure depicts the contribution of the false rejection rate (P_{FR}) and the false classification rate (P_{FC}) to the overall performance of the proposed system when automatically detected facial features were used for alignment and estimation of other locations for local feature extraction. **(A)** Curves for the ChimpZoo dataset. **(B)** Results for the ChimpTaï dataset. The black solid line denotes the line of equal error. The blue area represents the influence of the false rejections, while the red area shows the contribution of the false classifications. For the ChimpZoo dataset, the error rate of the system at the point of equal error is mainly caused by false rejection of genuine individuals, and only 3.52% is due to false classifications. For the ChimpTaï dataset, however, 13.52% of the overall error rate is caused by false rejections, while 17.72% are caused by false classifications.

based on the location of both eyes. For the manually annotated ones, however, we additionally used the annotated location of the mouth to estimate these locations more precisely. Therefore, the local feature extraction is much more accurate if an exact location of the mouth region is available.

In the previous three experiments, we showed that the face recognition algorithm proposed in this paper achieved excellent results and outperformed the approaches of previous works. However, we only showed the relationship between correct detection and identification rate (P_{DI}) and percentage of impostors accepted by the system (P_{FAR}). Moreover, another important question is how the system's overall error rate is influenced by the other two types of errors, the false rejection and the false classification. This issue is depicted in Figure 12, showing the results of the proposed system using automatically detected facial markings for ChimpZoo (Figure 12A) and ChimpTaï (Figure 12B). The blue area denotes the rate of false rejections (P_{FR}), while the red area shows the influence of false classification (P_{FC}) for different false alarm rates (P_{FA}). The lower bound depicts the ROC curve from Figure 11 (red dashed line). The false classification rates and false rejection rates for both datasets at the point of equal error can be seen in Table 5.

Table 5 P_{FR} and P_{FC} at EER for both datasets

	P_{FR} at EER	P_{FC} at EER
ChimZoo	0.1253	0.0352
ChimpTaï	0.1352	0.1773

False rejection rate (P_{FR}) and false classification rate (P_{FC}) at the point of equal error (EER) for both datasets, ChimpZoo and ChimpTaï, respectively.

It can be seen that for the ChimpZoo dataset, the main contribution to the overall error rate of the system is caused by falsely rejected faces of genuine individuals with P_{FR} of 12.53%. Only 3.52% was due to false classifications. This shows that many facial images of known identities were rejected as impostors because of too much pose variation, facial expressions, or occlusions. For the ChimpTaï dataset, however, the system's performance is almost equally caused by false classification, with P_{FC} of 17.73% and P_{FR} of 13.52%. This shows that the ChimpTaï dataset is much more challenging than the ChimpZoo dataset because it was gathered in a wildlife environment. Furthermore, the ChimpTaï dataset contains twice as much individuals at a much lower quality which again explains the strong influence of false classifications to the overall error of the proposed system.

5 Conclusions

In the ongoing biodiversity crisis, many species including great apes like chimpanzees for instance are threatened and need to be protected. An essential part of efficient biodiversity and wildlife conversation management is population monitoring and individual identification to estimate population sizes, asses viability, and evaluate the success of implemented protection schemes. Therefore, the development of new monitoring techniques using autonomous recording devices is currently of intense research [47]. However, manually processing large amounts of data is a tedious work and therefore extremely time-consuming and highly cost-intensive.

To overcome these issues, we presented an automated identification framework for chimpanzees in real-world

environments in this paper. Based on the assumption that humans and chimpanzees share similar properties of the face, we proposed to use the face detection and recognition technology for identification of great apes in our previous work [5-9]. In this paper we successfully combined face detection, face alignment, and face recognition to a complete identification system for chimpanzee faces in real-world environments. We successfully combined globally extracted holistic features and local descriptors for identification using a decision fusion scheme. As global features we used the well-established Gabor features. We transformed the resulting high-dimensional feature vectors into a smaller, more discriminating subspace using LPP. For classification we used an algorithm called SRC. Since it is known from the literature that different features encode different information, we also extract SURF around local facial feature points to make the system more robust against difficult lighting situations, various poses and expressions as well as partial occlusion by branches, leafs, or other individuals. We separately transformed the resulting SURF descriptors into a lower dimensional subspace for every facial fiducial point. After concatenating the resulting low-dimensional descriptors to get one comprehensive vector of local features, we use SVM with RBF kernel for classification. We combine the classification results of global and local features in a decision-based manner by taking the confidences of both classifiers into account. Furthermore, we thoroughly evaluated our proposed algorithm on two datasets of captive and free-living chimpanzee individuals which were annotated by experts using an open-set classification scheme. In the three experiments we showed that our approach outperforms previously presented algorithms for chimpanzee identification. Although both datasets were gathered in real-world environments, opposed to most datasets used to evaluate algorithms for human face recognition, our system performs very well and achieves promising results. Therefore, the presented framework can be applied in real-life scenarios for identification of great apes. Thus, the system will assist biologists, researchers, and gamekeepers with tedious annotation work of gathered image and video material and therefore has the potential to open up new venues for efficient and innovative wildlife monitoring and biodiversity conservation management. Currently, intensive pilot studies using autonomous infrared-triggered remote video cameras are conducted in Loango National Park, Gabon [48] and Taï National Park, Côte d'Ivoire [49]. These studies have provided promising results in both number of species detected, as well as visitation rates, demonstrating the potential of such an approach for biomonitoring. Our proposed framework for automatic detection and identification of chimpanzees will help researchers to efficiently scan and retrieve video sequences that are important for biologists, i.e., where chimpanzees or other great apes are present. After providing an annotated dataset of labeled chimpanzee faces, the system will also be able to recognize known and reject unknown individuals. Although grouping similar-looking faces of unknown individuals remains a future work, such an approach could help biologists to expand the dataset of known chimpanzees over time and successively improve the accuracy of the system. Hence, biologists are then able to conduct biodiversity time series analysis to assess whether significant fluctuations in biodiversity occur.

Although the presented system achieved very good results on both datasets, we hope to further increase the performance of the system by extending the approach for face recognition in video. Because the temporal component of video can contain important information for identification, we expect further improvement of the system by exploiting temporal information. For example, finding the shots in a video sequence which are best suitable for face recognition in terms of pose, motion blur, and lighting could be one approach to extend the system towards video. Furthermore, frame weighting algorithms or techniques like super-resolution are conceivable to take advantage of temporal information in video sequences. In addition, automatic detection of more facial features could lead to better alignment and more precise localization of facial fiducial points for local feature extraction, which will further improve the performance of the system.

Competing interests

Both authors declare that they have no competing interests.

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (BMBF) under the 'Pact for research and innovation'. We thank the Ivorian authorities for the long-term support, especially the Ivorian Ministére de l'Environnement, des Eaux et Forêts and the Ministére de l'Enseignement Supérieur et de la Recherche Scientifique, the directorship of the Taï National Park, the OIPR, and the CSRS in Abidjan. Financial support is gratefully acknowledged from the Swiss Science Foundation. We would like to thank especially Dr. Tobias Deschner for collecting videos and pictures over the last years and for providing invaluable assistance during the data collection. We thank all the numerous field assistants and students for their work on the Taï Chimpanzee Project. We thank the Zoo Leipzig and the Wolfgang Köhler Primate Research Center (WKPRC), especially Josep Call and all the numerous research assistants, zoo-keepers, and Josefine Kalbitz for their support and collaboration. We also thank Laura Aporius for providing the videos and pictures in 2010. This work was supported by the Max Planck Society. We also thank Laura Aporius for the annotation of data.

Author details

[1] Audio-Visual Systems, Fraunhofer IDMT, 98693 Ilmenau, Germany. [2] Electronic Imaging, Fraunhofer IIS, 91058 Erlangen, Germany.

Received: 31 January 2013 Accepted: 31 July 2013
Published: 19 August 2013

References

1. C Hilton-Taylor, SN Stuart, *Wildlife in a Changing World - an Analysis of the 2008 IUCN Red List of Threatened Species*, (Gland Switzerland, IUCN 2009). http://www.iucnredlist.org/technical-documents/references

2. PD Walsh, KAA Abernethy, M Bermejo, Catastrophic ape decline in western equatorial Africa. Nature. **422**, 611–614 (2003)
3. G Campbell, H Kuehl, PN Kouame, C Boesch, Alarming decline of west African chimpanzees in Côte d'Ivoire. Current Biology. **18**(19), R904–905 (2008)
4. JM Rowcliffe, C Carbone, Surveys using camera traps: are we looking to a brighter future? Anim. Conserv. **11**(3), 185–186 (2008)
5. A Loos, M Pfitzer, L Aporius, in *19th European Signal Processing Conference (EUSIPCO)*. Identification of great apes using face recognition (Barcelona, 29 August–2 September 2011)
6. A Loos, in *1st ACM International Workshop on Multimedia Analysis for Ecological Data (MAED) in Conjunction with ACM Multimedia*. Identification of great apes using Gabor features and locality preserving projections (ACM, New York, 2012)
7. A Loos, A Ernst, in *IEEE International Symposium on Multimedia*. Detection and identification of chimpanzee faces in the wild (Irvine, CA, 10–12 December 2012)
8. A Loos, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Identification of primates using global and local features (Vancouver, 26–31 May 2013)
9. A Ernst, C Küblbeck, Fast face detection and species classification of African great apes, *8th IEEE International Conference on Advanced Video and Signal Based Surveillance*. (IEEE, New York, 2011), pp. 279–284
10. H Bay, A Ess, T Tuytelaars, LV Gool, Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
11. HA Rowley, S Baluja, T Kanade, Neural network-based face detection. IEEE Trans. Pattern Anal. Mach. Intell. **20**, 23–38 (1998)
12. P Viola, M Jones, Rapid object detection using a boosted cascade of simple features, *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1 (IEEE, New York, 2001), pp. 511–518
13. Y Freund, RE Schapire, A short introduction to boosting. J. Japanese Soc. Artif. Intell. **14**(5), 771–780 (1999)
14. P Viola, M Jones, Robust real-time object detection. Int. J. Comput. Vis. **57**(2), 137–154 (2002)
15. R Lienhart, J Maydt, An extended set of haar-like features for rapid object detection, *IEEE International Conference on Image Processing (ICIP)*, vol. 1. (IEEE, New York, 2002), pp. 900–903
16. B Wu, A Haizhou, H Chang, L Shihong, Fast rotation invariant multi-view face detection based on real Adaboost, *6th IEEE International Conference on Automatic Face and Gesture Recognition* (IEEE, New York, 2004), pp. 79–84
17. J Wawerla, S Marshall, G Mori, K Rothley, P Sabzmeydani, BearCam: automated wildlife monitoring at the Arctic Circle. Mach. Vis. Appl. **20**(5), 303–317 (2009)
18. T Burghardt, J Calic, in *8th Seminar on Neural Network Applications in Electrical Engineering*. Real-time face detection and tracking of animals (Serbia & Montenegro, Belgrade, 25–27 September 2006), pp. 27–32
19. C Spampinato, D Giordano, R Di Salvo, in *ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams (ARTEMIS)*. Automatic fish classification for underwater species behavior understanding (Firenze, Italy, 29 October 2010), pp. 40–50
20. C Spampinato, S Palazzo, B Boom, J van Ossenbruggen, I Kavasidis, R Di Salvo, FP Lin, D Giordano, L Hardman, RB Fisher, Understanding fish behavior during typhoon events in real-life underwater environments. Multimedia Tools Appl (2012). doi:10.1007/s11042-012-1101-5. http://scholar.google.com/citations?view_op=view_citation&hl=de&user=yJr6TqAAAAAJ&citation_for_view=yJr6TqAAAAAJ:qjMakFHDy7sC
21. M Turk, A Pentland, Eigenfaces for recognition. J Cogn. Neurosci. **3**, 71–86 (1991)
22. PN Belhumeur, JP Hespanha, DJ Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Trans. Pattern Anal. Pattern Anal. Mach. Intell. **19**(7), 711–720 (1997)
23. X He, S Yan, Y Hu, P Niyogi, HJ Zhang, Face recognition using Laplacianfaces. IEEE Trans. Pattern Anal. Mach. Intell. **27**(3), 328–40 (2005)
24. J Wright, AY Yang, A Ganesh, SS Sastry, Y Ma, Robust face recognition via sparse representation. IEEE Trans. Parallel Patt. Anal. Mach. Intell. **31**(2), 210–27 (2009)
25. M Yang, L Zhang, in *European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science*. Gabor feature based sparse representation for face recognition with Gabor occlusion dictionary, vol. 6316 (Springer, Heidelberg, 2010), pp. 448–461
26. A Ardovini, L Cinque, E Sangineto, Identifying elephant photos by multi-curve matching. Pattern Recognit. **41**(6), 1867–1877 (2008)
27. BN Araabi, N Kehtarnavaz, T McKinney, G Hillman, B Würsig, A string matching computer-assisted system for dolphin photoidentification. Ann. Biomed. Eng. **28**(10), 1269–1279 (2000)
28. T Burghardt, N Campbell, in *5th International Conference on Computer Vision Systems (ICVS)*. Individual animal identification using visual biometrics on deformable coat patterns (Bielefeld, 21–24 March 2007). http://biecoll.ub.uni-bielefeld.de/volltexte/2007/20/
29. T Burghardt, Visual animal biometrics - automatic detection and individual identification by coat pattern, PhD Thesis, University of Bristol, 2008.
30. M Lahiri, R Warungu, DI Rubenstein, TY Berger-Wolf, C Tantipathananandh, Biometric animal databases from field photographs: identification of individual zebra in the wild, *ACM International Conference on Multimedia Retrieval (ICMR)* (ACM, New York, 2011)
31. RE Schapire, Y Singer, Improved boosting algorithms using confidence-rated predictions. Mach. Learn. **37**(3), 297–336 (1999)
32. B Fröba, *Verfahren zur Echtzeit-Gesichtsdetektion in Grauwertbildern*. (Shaker, Aachen, 2003)
33. R Zabih, J Woodfill, in *Third European Conference on Computer Vision Proceedings, Volume II*. Non-parametric local transforms for computing visual correspondence (Stockholm, Sweden, 2–6 May 1994), pp. 151–158
34. T Ojala, M Pietikäinen, D Harwood, A comparative study of texture measures with classification based on featured distributions. Pattern Recognit. **29**, 51–59 (1996)
35. Y Gao, Y Wang, X Feng, X Zhou, Face recognition using most discriminative local and global features. *International Conference on Pattern Recognition (ICPR)* (IEEE, New York, 2006), pp. 351–354
36. L Wiskott, JM Fellous, N Krüger, C von der Malsburg, Face recognition by elastic bunch graph matching. IEEE Trans. Pattern Anal. Mach. Intell. **19**(7), 775–779 (1997)
37. C Liu, H Wechsler, Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans. Image Proc. **11**(4), 467–476 (2002)
38. S Xie, S Shan, X Chen, Fusing local patterns of Gabor magnitude and phase for face recognition. IEEE Trans. Image Proc. **19**(5) (2010)
39. D Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004)
40. CC Chang, CJ Lin, LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Tech. **2** (2011). doi:10.1145/1961189.1961199. http://dl.acm.org/citation.cfm?id=1961199
41. B Gökberk, AA Salah, L Akarun, Rank-based decision fusion for 3D shape-based face recognition. *International Conference on Audio- and Video-Based Biometric Person Identification (AVBPA)* (Springer-Verlag, Berlin Heidelberg, 2005), pp. 1019–1028
42. TF Wu, CJ Lin, R Weng, Probability estimates for multi-class classification by pairwise coupling. J Mach. Learn. Res. **5**, 975–1005 (2004)
43. A Ernst, T Ruf, C Küblbeck, in *Proceedings of the 2nd Workshop on Pervasive Advertising 2009 In conjunction with Informatik 2009*. A modular framework to detect analyze faces for audience measurement systems (Lübeck, Germany, 2 October 2009), pp. 3941–3953
44. PJ Phillips, S Rizvi, P Rauss, The FERET evaluation methodology for face recognition algorithms. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 1090–1104 (2000)
45. PJ Phillips, P Grother, R Micheals, D Blackburn, E Tabassi, J Bone, Face recognition vendor test 2002: evaluation report. (Nistir 6965, National Institute of Standards and Technology, 2003)
46. PJ Phillips, P Grother, R Micheals, in *Handbook of Face Recognition*. Chapter 21: evaluation methods in face recognition (Springer-Verlag, London, 2011), pp. 551–574
47. JA Ahumada, CEF Silva, K Gajapersad, C Hallam, J Hurtado, E Martin, A McWilliam, B Mugerwa, T O'Brien, F Rovero, D Sheil, WR Spironello, N Winarni, SJ Andelman, Community structure and diversity of tropical forest mammals: data from a global camera trap network. Philos. Trans. R. Soc. London B. **366**(1578), 2703–2711 (2011)

48. J Head, C Boesch, M Robbins, L Rabal, L Makaga, H Kuehl, Effective socio-demographic population assessment of elusive species for ecology and conservation management. Ecol. Evol (2013). doi:10.1002/ece3.670. http://onlinelibrary.wiley.com/doi/10.1002/ece3.670/abstract
49. B Hoppe-Dominik, H Kühl, G Radl, F Fischer, Long-term monitoring of large rainforest mammals in the biosphere reserve of Taï National Park, Côte d'Ivoire. Afr. J. Ecol. **49**(4), 450–458 (2011)

doi:10.1186/1687-5281-2013-49
Cite this article as: Loos and Ernst: **An automated chimpanzee identification system using face detection and recognition.** *EURASIP Journal on Image and Video Processing* 2013 **2013**:49.