Classification of Bird Species from Video Using Appearance and Motion Features

John Atanbori[a,*], Wenting Duan[b], Edward Shaw[d], Kofi Appiah[c], Patrick Dickinson[b]

[a] *University of Nottingham, School of Computer Science, Nottingham, UK*
[b] *University of Lincoln, School of Computer Science, Lincoln, UK*
[c] *Sheffield Hallam University, Department of Computing, Sheffield, UK*
[d] *Don Catchment Rivers Trust, St Catherines House, Woodfield Park, Doncaster, UK*

Abstract

The monitoring of bird populations can provide important information on the state of sensitive ecosystems; however, the manual collection of reliable population data is labour-intensive, time-consuming, and potentially error prone. Automated monitoring using computer vision is therefore an attractive proposition, which could facilitate the collection of detailed data on a much larger scale than is currently possible.

A number of existing algorithms are able to classify bird species from individual high quality detailed images often using manual inputs (such as a priori parts labelling). However, deployment in the field necessitates fully automated in-flight classification, which remains an open challenge due to poor image quality, high and rapid variation in pose, and similar appearance of some species. We address this as a fine-grained classification problem, and have collected a video dataset of thirteen bird classes (ten species and another with three colour variants) for training and evaluation. We present our proposed algorithm, which selects effective features from a large pool of appearance and motion features. We compare our method to others which use appearance features only, including image classification using state-of-the-art Deep Convolutional Neural Networks (CNNs). Using our algorithm we achieved an 90% correct classification rate, and we also show that using effectively selected motion and appearance features together can produce results which outperform state-of-the-art single image classifiers. We also show that the most significant motion features improve correct classification rates by 7% compared to using appearance features alone.

Keywords: Appearance Features, Motion Features, Feature Extraction, Feature Selection, Bird Species Classification, Fine-Grained Classification

1. Introduction

Fine-grained object classification is a well know challenge in computer vision, and there has been some previous work on bird species recognition using individual images (Gavves et al., 2015; Berg et al., 2014; Berg and Belhumeur, 2013; Huang et al., 2013; Branson et al., 2014; Duan et al., 2012). Robust automated classification would be of potential benefit to ecologists studying bird populations; however, there are significant limitations with existing works, which are trained and tested on high quality images. Such images are difficult to capture in real-world settings. In addition, most existing methods are not fully automated, limiting the data which can be processed. For useful deployment, such systems need to classify birds in flight, and this has also not yet been fully studied. In-flight classification introduces challenges around image quality, shape, and image noise, but also presents some opportu-

*Corresponding author:
Email address: john.atanbori@nottingham.ac.uk (John Atanbori)

nities: the flight patterns of birds are known to vary across different species (Bruderer et al., 2010; Duberstein et al., 2012) and are used by human observers to assist recognition. Very few existing studies (Cullinan et al., 2015; Matzner et al., 2015) have made use of motion features in this way, and these have only differentiated small numbers of species. None has previously combined appearance and motion features to facilitate improved classification.

We focus on this challenge, and present our system which can reliably classify thirteen bird classes in flight. In our previous work (Atanbori et al., 2016), we presented separate sets of appearance and motion features, and showed that our appearance features out-performed the state-of-the-art (Marini et al., 2013). We further showed in Atanbori et al. (2015) that motion features can be used for real-time classification. In this paper we present the following new contributions:

1. We combine appearance and motion features to classify bird species in flight.
2. We introduce an extended dataset, which contains video data of thirteen bird classes. This presents a significant challenge, and is representative of real-world monitoring contexts.
3. We compare feature selection methods, with four standard classifiers, Random Forest (RF), Random Tree (RT), Naive Bayes (NB), and Support Vector Machines (SVM), and compare performance.
4. We compare our method with two state-of-the-art deep learning models (VGG19 and MobileNet), which are used to classify individual images from our dataset.
5. We have demonstrated that the most significant motion features improve correct classification rate compared to using appearance features only.

The remainder of this paper is structured as follows. In Section 2, we review existing work, including applications to other species. In Section 3, we introduce our dataset and processing architecture, then proceed to Section 4 which describes our motion and appearance features. In Sections 4, 5 and 6 we describe our feature selection, classifiers, and experimental setup, and conclude in Sections 7 and 8, by presenting our evaluation and results.

2. Related Work

A common approach taken by many computer vision researchers in monitoring and identification of animals is to exploit the feature selection techniques, e.g. (Matzner et al., 2015; Spampinato et al., 2010; Beyan, 2015). Hence, in this section we review relevant papers from this domain, and conclude with a review of feature selection techniques.

2.1. Studies related to other Species

Colour features are generally undetectable in low-light, and so applications to bat monitoring have looked more closely at motion features. For example, Cullinan et al. (2015); Matzner et al. (2015); Hristov et al. (2010); Betke et al. (2008); Lazarevic et al. (2008) censused large bat populations. Betke et al. (2008) also estimated wing beat frequencies of individuals using pose templates, and applying Fast Fourier Transform (FFT). Previous work by ourselves, Atanbori et al. (2013), also used FFT to determine wing beat frequency, but using a bounding box fitted to the segmented silhouette. Like Betke et al., works by Cullinan et al., Matzner et al., and Lazarevic et al. use thermal imaging. Cullinan et al. defined four classes (bat, gull, tern, swallow), and using flight tracks reported 82% correct classification; however, this does not consider fine-grained differentiation.

Lee et al. (2003) used shape contour features to discriminate between nine fish species, and achieved a classification rate of between 13% and 80%. Spampinato et al. (2010) also used texture and shape features, achieving a 92% correct rate. Rodrigues et al. (2010) used Scale-Invariant Feature Transform (SIFT) and Principal Component Analysis (PCA), achieving similar results to Lee et al., and Spampinato et al. (92% across six species).

2.2. Classification of Bird Species

Whilst Cullinan et al. (2015), and Matzner et al. (2015) used motion features to differentiate between small numbers of bird species, other existing works concerned with automated classification of birds use appearance features from a single image of an individual bird. These approaches can be further subdivided into those that make use of the physical structure of the bird (which we refer to as *part-based*), and those which do not. Non-part-based

methods use colour and shape features, without considering their relative position or orientation (Marini et al., 2013; Wah et al., 2011a,b). For example, Marini et al. uses colour features with SVM. Again, these have been used to differentiate between small numbers of species, and struggle to maintain performance as the number increases. Marini et al. showed that when using colour features alone on the Caltech-ucsd birds-200-2011 dataset, accuracy reduces from approximately 85% when selecting between 2 species to 20% when differentiating between 17 species.

Part-based methods associate features with specific body parts (Wah et al., 2011a; Duan et al., 2012; Berg and Belhumeur, 2013; Huang et al., 2013; Branson et al., 2014; Wah et al., 2011b; Berg et al., 2014). This can help differentiate between species with high visual correlation, but almost all require manual inputs, and good-quality images. Berg et al. developed an online application called Birdsnap; this requires manual annotation of parts prior to segmentation and classification. Krause et al. (2015), and Gavves et al. (2013, 2015) both developed annotation-free parts-based methods. Their results compared favourably: based on CUB-2011, correct classification rates were 62%, 82% and 67% respectively. They used the Grab-Cut segmentation method (Rother et al., 2004), and so still require some manual intervention. Another annotation free method was proposed by Zhang et al. (2015), which detects parts using Convolutional Neural Network (CNN). Whereas Krause et al. align the co-segmented objects before labelling, Zhang et al. (2015) uses CNN feature to detects parts automatically, but did not improve significantly on other methods.

Our objective is to develop a deployable system, capable of identifying birds in flight. Existing methods require manual annotation and/or high quality images which exceed the quality available in real-world settings. In addition, we assert that birds far from the camera are less easily classified using appearance features alone. This motivates our approach of combining colour and motion features. Of the other approaches mentioned, we consider that Marini et al. is an appropriate comparator for this problem domain, being fully automated, non-parts based (so more likely to be robust to reduced image quality), and reporting good results. We therefore use this method as an initial benchmark for our work. Deep learning is the current state-of-the-art in image classification; we have thus also used two recent CNN models as additional benchmarks.

2.3. CNN Classification Methods

As mentioned in Section 2.2, Krause et al. and Zhang et al. (2015) used CNN methods for single image species classification. CNN has become prevalent in computer vision for image classification since (Krizhevsky et al., 2012) won the ImageNet Challenge in 2012. Deep learning has been used to achieve state-of-the-art accuracy on ImageNet (Deng et al., 2009), and the PASCAL VOC (Everingham et al., 2012) datasets. Important recent architectures include VGG16 and VGG19 (Simonyan and Zisserman, 2014), which are very deep networks achieving state-of-the-art classification on the 2014 ImageNet challenge. The original VGG19 model comprised 144 million parameters; however, MobileNets (Howard et al., 2017) is another state-of-the-art deep learning model that uses depth-wise separable and point-wise convolutions to reduce the number of model parameters (to only 4.2 million). This model trades accuracy against resources, but still achieved classification accuracy comparable to very deep networks with many large numbers of parameters. We have evaluated our work against both state-of-the-art networks (VGG19 and MobileNet) and presented results in section 7.

2.4. Feature Selection and Reduction

We make extensive use of feature selection and reduction in the method we present in this paper, and so include a short review of relevant techniques here. Redundant or irrelevant features may affect classification rates negatively (Hall, 1999; Yu and Liu, 2003). Thus, feature reduction may be important for fine-grained identification. Existing feature selection methods can broadly be divided into *filter* and *wrapper* methods.

Filter methods rank the significance of proposed features. A number of ranking criteria have been used including Fisher score (Gu et al., 2012), Pearson correlation coefficient (Guyon and Elisseeff, 2003), mutual information (Peng et al., 2005; Yu and Liu, 2003) and Relief (Robnik-Šikonja and Kononenko, 2003; Moore and White, 2007). Filter methods are efficient (Lee et al., 2012), and scale well.

Wrapper methods use machine learning to evaluate the effectiveness of feature subsets (Tang et al., 2014). An example was proposed by Breiman (2001), and is based on Variable Importance (VI) derived from Classification and Regression Trees (CART), and Random Forests. However, if the data contains groups of correlated features of similar importance, then smaller groups are favoured (Toloşi and Lengauer, 2011). Hall (1999) and Hall et al. (2009) proposed some methods to combine the filter and wrapper methods. Hall reported the performance and accuracy of this approach to be better than wrappers on some datasets.

3. Dataset and Preprocessing

Figures 1, 2 and 3 show samples from our video dataset, recorded using a Casio Exilim ZR100 camera at 240 frames per second. There are thirteen classes in total, made up of ten unique species and another (*Melopsittacus Undulatus*) with three colour variants. The dataset is made up of videos recorded over several days, from three different sites and each class is made up of at least ten individuals. Table 1 shows the number of videos and images per species. The majority (most represented) class is the Black-headed Gull (*Chroicocephalus ridibundus*), and the minority is the Common wood pigeon (*Columba palumbus*).

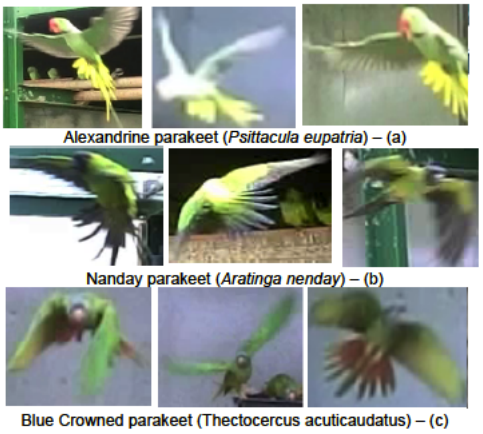

Figure 1: **Parakeet samples from our dataset**. **a)** Alexandrine parakeet (*Psittacula eupatria*), **b)** Nanday parakeet(*Aratinga nenday*) and **c)** Blue-crowned parakeet (*Thectocercus acuticaudatus*)

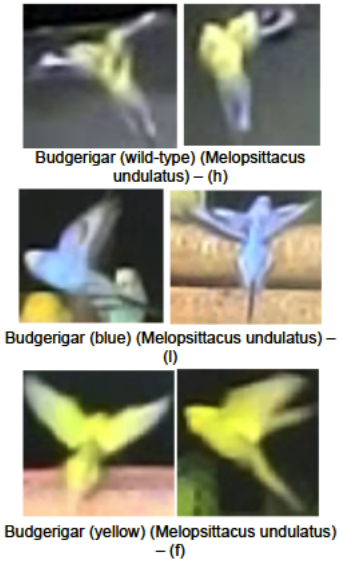

Figure 2: **Budgerigar samples from our dataset**. **f)** Budgerigar (yellow) (*Melopsittacus undulatus*), **h)** Budgerigar (wild-type) (*Melopsittacus Undulatus*) and **l)** Budgerigar (blue) (*Melopsittacus Undulatus*)

For each frame in each video, we automatically extracted the target bird silhouette (Fig. 4) using a background Gaussian Mixture Model (GMM) (Zivkovic and van der Heijden (2006)). We used the method proposed by Suzuki et al. (1985) to obtain contours and oriented bounding boxes. A selection of metrics (height, width and hypotenuse, centroid, silhouette and contour points) was automatically determined. We also extracted and concatenated colour moments, shape moments, greyscale histogram, Gabor filter and log-polar features (see Section 4). For motion features, we formed a trajectory with the 2D centroid position, described as:

$$T = \{(x_1, y_1), \ldots, \left(x_j, y_j\right), \ldots, (x_N, y_N)\} \qquad (1)$$

Where N is the number of frames in T, represented as a series of x and y coordinates. We segmented the trajectory (Equation 1) into overlapping sub-trajectories t_k of equal length, where $k = 1 \ldots N - Q + 1$ is the total number of sub-trajectories:

$$t_k = \{(x_k, y_k), \ldots, (x_{k+Q-1}, y_{k+Q-1})\} \qquad (2)$$

We used $Q = 64$, where $Q \leq N$, and applied a box

Figure 3: **Other samples from our dataset**. **d)** Common house martin (*Delichon urbica*), **e)** Eastern Rosella (*Platycercus eximius*), **g)** House sparrow (*Passer domesticus*), **i)** Common wood pigeon (*Columba palumbus*), **j)** Black-headed gull (*Chroicocephalus ridibundus*), **k)** Cockatiel (*Nymphicus hollandicus*), and **m)** Common starling (*Sturnus vulgaris*)

filter (Gonzalez and Woods, 2002) with a 1 x 3 kernel to reduce noise. We extracted motion features from the set of smoothed short trajectories (see Equation 2).

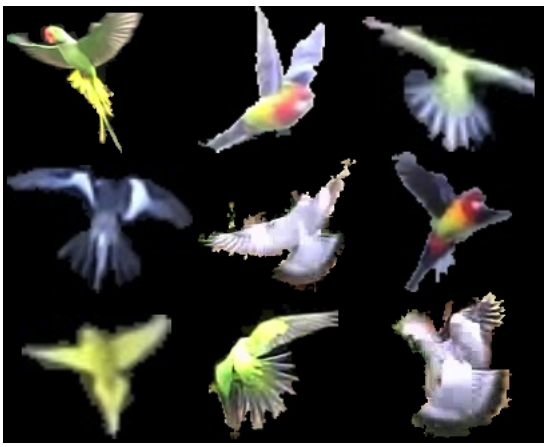

Figure 4: Flying birds segmented from our data.

Table 1: Number of videos and images of the thirteen classes

Species	# of videos	# of images	# of images/total(%)
j = Black Headed Gull	147	38,764	24%
d=Common House Martin	139	25,517	16%
a = Alexandrine Parakeet	79	12,801	8%
l=Budgerigar (blue)	81	12,090	7%
g =House Sparrow	78	10,191	6%
b = Nanday Parakeet	60	10,025	6%
m=Common Starling	71	9,865	6%
k = Cockatiels	59	9,398	6%
c=Blue-crowned Parakeet	60	9,076	6%
f=Budgerigar (yellow)	54	7,667	5%
h=Budgerigar (wild-type)	48	6,283	4%
e=Eastern Rosella	44	5,929	4%
i=Common Wood Pigeon	37	4,301	3%
Total	**957**	**161,907**	**100%**

3.1. Additional Preprocessing for CNN Classifiers

For our benchmark comparisons, we used the VGG19 (Simonyan and Zisserman, 2014) and MobileNet (Howard et al., 2017) deep learning architectures: these were used to classify individual image frames from the video datasets. For each video frame, we fitted abounding

box around the tracked bird's silhouette. We rescaled the silhouette to a standard size for training and testing. In our case, we used 32x32 image sections, as the majority of the silhouettes in the database are of around this size.

4. Feature Extraction

In this section, we describe a pool of appearance and motion features extracted to form our method, which we did not use in the benchmark methods.

4.1. Appearance Features

Appearance features are collated from existing related works and comprise colour moments and log-polar values, shape moments, Gabor filters and greyscale histograms. We have previously shown in Atanbori et al. (2016) that this feature set is useful for bird classification.

4.1.1. Colour Moment Features

Histogram features have been used widely to characterise colour images, by constructing histograms across multiple colour channels (Sergyan, 2008; Huang et al., 2010). Statistical features have also been extracted from such histograms and used to describe colour-based features of bird species. We compute colour moment features from each bird silhouette by first transforming the colour image from RGB to HSV space, and then building a histogram to represent the distribution of values across each channel. We compute colour moment features by transforming the image from RGB to HSV and forming a histogram of colour values. We use 30 bins for the Hue channel, and 32 each for Saturation and Value, then normalise and calculate statistics for each (Mean, Standard Deviation, Skewness, Energy and Entropy, see Sergyan (2008); Huang et al. (2010)). Each bin value and statistic is used as a single-valued feature, providing 109 in total.

4.1.2. Image Moment Features

Image moments are used in computer vision to find the area (or total intensity) of the segmented object including information about its centroid and orientation. We used spatial moments (Jacob et al., 2001) and Hu moments (Du et al., 2007), to represent the shape of the silhouette in each frame. We first extract the contours (Suzuki and Abe, 1985) followed by the seven Hu moments and ten spatial moments

4.1.3. Greyscale Histogram Features

Greyscale Histogram Features are constructed by first converting the extracted silhouette into a greyscale image. A 256-bin histogram is then created from the greyscale silhouette (excluding background), to form a representation of the greyscale distribution of pixels. We then compute statistics from the bins: mean, standard deviation, skewness, kurtosis, energy, entropy, and Hu's 2nd and 3rd moments, providing eight features.

4.1.4. Gabor Wavelet Features

Gabor wavelet features have both multi-resolution and multi-orientation properties, are optimal for measuring local spatial frequencies and yield distortion tolerance space for pattern recognition tasks. The Gabor wavelet transform (Lee, 1996) is the convolution between the function g and image $I(x, y)$, given by Equation 3.

$$g = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right)\exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \qquad (3)$$

where $x' = x\cos\theta + y\sin\theta$, $y' = -x\sin\theta + y\cos\theta$ and θ , λ, ψ, γ and σ are orientation, wavelength, phase, aspect ratio and standard deviation respectively.

We extract these, using $\lambda = 1$, $\psi = 0$, $\gamma = 0.02$, $\sigma = 1$, and Scale = 31. We use four different values for $\theta = \{0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}\}$. The result is four greyscale wavelet images, from which we extract mean, standard deviation, skewness, kurtosis, energy, entropy, yielding 20 features.

4.1.5. Log-Polar Features

Pun and Lee (2003) demonstrated that log-polar images can eliminate transformation effects. An image I is transformed into a log-polar form $dst(\theta,\rho)$ using Equation 4.

$$dst(\theta,\rho) \leftarrow src(x,y)\, for \begin{cases} \rho = \log\sqrt{x^2 + y^2} \\ \theta = \arctan\left(\frac{x}{y}\right) if\, x > 0 \end{cases} \qquad (4)$$

We convert the segmented image to HSV space and apply a log-polar transform to each channel. We then compute mean, standard deviation, skewness, entropy and energy, for each.

4.2. Motion Features

We extracted motion features using a time window described by Equation 2 and detailed these features in this section.

4.2.1. Curvature Scale Space (CSS)

Both Beyan and Fisher (2013) and Mai et al. (2010) used Curvature Scale Space (CSS) to distinguish trajectories; these are robust to noise (Mai et al., 2010), and calculated using Equation 5.

$$K_i = \frac{x_i^{'} y_i^{''} - y_i^{'} x_i^{''}}{\left(x_i^{'2} + y_i^{'2}\right)^{\frac{3}{2}}} \tag{5}$$

where $x_i^{'}$, $x_i^{''}$, $y_i^{'}$ and $y_i^{''}$ are first and second derivatives of x_i and y_i respectively.

We formed 22 CSS features which includes ten statistics (mean, standard deviation, skewness, kurtosis, entropy, minimum, maximum, local minima, local maxima and zero crossings) from the absolute curvature, the number of curves in the CSS image, the total length of all the curves and ten statistics computed from the CSS maxima (as for the absolute curvature). The CSS feature was considered important for two reasons: Firstly, the datasets consist of birds flying in different directions and orientations. Therefore, similar trajectories may appear as a different relative orientation with respect to the optical axis. Secondly, they may appear at different scales due to varying distances from the camera.

4.2.2. Turn Based Features

Trajectory turnings were used to represent the shape of flight, and are calculated by computing the slope of the trajectory between consecutive frames. They have been used, similar to those in Li et al. (2006) and Beyan and Fisher (2013) and were computed using equation 6:

$$\theta_k = \begin{cases} arctan(\frac{y_k - y_{k-1}}{x_k - x_{k-1}}), \\ \text{if } (x_k - x_{k-1}) > 0. \\ \\ arctan(\frac{y_k - y_{k-1}}{x_k - x_{k-1}}) + \pi, \\ \text{if } (x_k - x_{k-1}) \leq 0, (y_k - y_{k-1}) \geq 0. \\ \\ arctan(\frac{y_k - y_{k-1}}{x_k - x_{k-1}}) - \pi, \\ \text{if } (x_k - x_{k-1}) \leq 0, (y_k - y_{k-1}) < 0. \end{cases} \tag{6}$$

Where $(x_k - x_{k-1})^2 - (y_k - y_{k-1})^2 \neq 0$. A histogram (with bin of size 3) was used to calculate the ten statistical features, forming a subset of 30 features.

4.2.3. Wing Beat Frequency Features

The periodicity of wings beats is known to vary among bird species (Lazarevic et al., 2008). In Atanbori et al. (2013) we show that for bat species, a bounding box tted to the silhouette of a tracked individual can be used to measure the periodicity of wing beats. We used the same approach in this work and extracted three metrics (height, width and diagonal length) from a bounding box fitted to the bird silhouette. For each frame, we then computed the frequency components of the signal using a sequence of values computed from a time window of N frames, centred on the current frame. Each frequency component is computed using the Fast Fourier Transform (FFT) 7, for each metric separately:

$$F(k) = \sum_{t=0}^{(N-1)} f(t)\, e^{-i2\pi kt/N} \tag{7}$$

Where $f(t)$ is the signal in the spatial domain, t ranges from $0 \ldots N-1$, and N is the number of frames in the time window. $F(k)$ is the k^{th} frequency domain component (encoding both amplitude and phase) with $k = 0 \ldots N-1$. In our case, we used $N = 64$ frames but added padding of 32 zeroes before and after, to produce a total sequence length of$N = 128$. The zero padding is added to increase the resolution of the computed frequencies (as this may help discriminate species with closely related frequencies). This value of N was appropriate for our dataset, but for longer sampled trajectories, large values of N could

be used to increase resolution. Having generated the frequency components, the nine most dominant frequencies for each metric (excluding the DC component) were used to form a set of 27 features for the frame.

4.2.4. Centroid Distance Function (CDF)

The centroid distance function (CDF) represents trajectory shape (Beyan and Fisher, 2013) and is invariant to translation and rotation. Since the flying bird's trajectory are subject to rotational deformation, we used CDF features which we computed using Equation 8. We then extracted the ten statistical moments (as with the CSS features) to form this feature set.

$$CDF_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \tag{8}$$

where:

$$x_c = \frac{1}{N}\sum_{j=0}^{N-1} x_j, \qquad y_c = \frac{1}{N}\sum_{j=0}^{N-1} y_j \tag{9}$$

$i = 0, 1, ...N-1$, and N is the number of points.

4.2.5. Vicinity

We also included normalised *vicinity* features which were previously used in Liwicki et al. (2006), and calculate ten statistical moments from each (Vicinity curliness, slope, aspect and linearity) to form a subset of 40. Vicinity features were selected to represent part of the motion features since they consist of features extracted from each point and takes into consideration their neighbouring points and are very robust to noisy data.

4.2.6. Curvature

Birds flight have directional bearings, which can be measured between frames using curvature. The curvature is computed as the cosine of the angle between the line from a point to its predecessor, and the following line. Given a trajectory t_k and successive points $P_{k-1}, P_k and P_{k+1}$, the curvature $\cos(\theta_k)$ is given by Equation 10.

$$\cos(\theta_k) = \frac{c^2 - a^2 - b^2}{2ab} \tag{10}$$

Where: a is the distance from trajectory point $P_{k-1}(x_{k-1}, y_{k-1})$ to $P_k(x_k, y_k)$ given by $\left((x_{k-1} - x_k)^2 + (y_{k-1} - y_k)^2\right)^{0.5}$. Similarly, b is given by $\left((x_k - x_{k+1})^2 + (y_k - y_{k+1})^2\right)^{0.5}$ and c is given by $\left((x_{k-1} - x_{k+1})^2 + (y_{k-1} - y_{k+1})^2\right)^{0.5}$. $k = 2 \ldots Q-1$ and $Q = 64$ is the length of the short trajectory defined in Section 3

Ten statistical features including mean, maximum, minimum, standard deviation, number of zero crossings, local minima and maxima, skewness, energy and entropy of the curvature were extracted.

5. Feature Selection

Our feature set, described in Section 4, totals 320. We hypothesise that some of these are redundant, and have investigated two selection strategies: correlation-based and classifier-based. We describe each below.

5.1. Correlation-based Feature Selection

The correlation-based method is based on Hall et al. (2009). We calculated a matrix of feature-to-class and feature-to-feature Pearsons correlations, and search the subset space (Hall, 1999) to determine the most effective. The *merits* of each feature subset are then computed using equation 11.

$$M_s = \frac{k\bar{r_{cf}}}{\sqrt{k + k(k-1)\bar{r_{ff}}}} \tag{11}$$

Where M_s is the merit of k features, $\bar{r_{cf}}$ is the mean correlation ($f \in S$) and $\bar{r_{ff}}$ is the average feature-to-feature inter-correlation. The feature-to-feature correlation can be expressed as:

$$\bar{r_{ff}} = \left(\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} r_{ff}\{f_i, f_j\}\right) / {}^{k}C_2 \tag{12}$$

Where $r_{ff\{f_i,f_j\}}$ is the pairwise correlation of feature f_i with f_j, and ${}^{k}C_2$ is the number of combinations possible from the subset S. The feature-class correlation is calculated as:

$$\bar{r_{fc}} = \left(\sum_{i=1}^{k} r_{fc}\{f_i, c_i\}\right) / k \tag{13}$$

Where $r_{fc\{f_i,c_i\}}$ is the pairwise correlation of feature f_i with class c_i.

5.2. Classifier-based Feature Selection Method

The classifier-based method uses a Random Forest classifier Breiman (2001) with "Bagging" (Breiman, 1996), Like Breiman (2001), we use permutation importance to calculate tree splits.

6. Experiments

In this section, we describe our experiments, including setup, methodology and benchmarks. For our evaluation we have performed three separate sets of experiments:

- We have quantified the effectiveness of our complete feature set across our dataset using four well-known classifiers.
- Evaluated the two feature selection methods, and have identified the most effective subsets from our pool of features.
- We have compared the results of our method with that of image classification using two deep learning networks (VGG19 and MobileNet).
- We demonstrated that most significant motion features contribute to classifier effectiveness.

6.1. Setup and Method

For experiments using our proposed method, we computed the full set of appearance and motion features for the frames in each video, for each class of bird. These features were concatenated to create our full feature set (total of 320 features:169 appearance and 151 motion features). We sampled individual image frames from the dataset for all and split the dataset into 80% for training and 20% for testing, which was used to evaluate the effectiveness of the features using four classifiers: the Naive Bayes (NB), Random Forest (RF), Random Tree (RT) and Support Vector Machine (SVM). The SVM classifier is base on LibSVM proposed by Chang and Lin (2011), which is comparable to the one used in Marini et al. (2013) and implemented using a radial basis function kernel. The gamma and cost parameters were optimised using a grid search. We used $K = int(log_2(\#features) + 1)$ randomly chosen attributes at each node for the RT classifier, with unlimited depth. Convergence of the *out of bag* errors for RF occurred at 20 trees. The NB classifier assumed a Gaussian mixture model over the whole training data distribution, one component per class, and estimated parameters from the training data. All our experiments were performed on a Mac book pro laptop running OS X 10.9.5, with 2.5 GHz Processor and 4GB RAM. We used C++ with XCode 5.1.1 and OpenCV 3.0 to implement all our pre-processing and feature extraction algorithms and WEKA 3.7 (Hall et al., 2009) for the classification and feature selection algorithms. The results are reported in Section 7.

To investigate performance after feature selection, the correlation-based merits M_s (see Equation 11) were sorted in descending order. Starting with the complete set of 320 features, we iteratively removed the ten least significant until only ten remained, and used the classification rate to plot a learning curve. We used a similar scheme for the classifier-based method and used the curves to identify the optimal parameters and reported the results in Section 7. We applied these techniques to our 169 appearance features only to help to compare our best appearance features to the full feature set and evaluate the contribution of our motion features to classification performance. We applied a cost matrix of pairwise class error weightings based on the degree of taxonomic relatedness of the bird classes (see Table 2) for both feature selection and classification. The cost matrix used taxonomic relatedness weightings of 0.25, 0.5, 0.75 and 1 for species, family, order and class respectively. We used the approach from Domingos (1999) as it is independent of the actual classifying technique that is used and has a similar implementation in Weka. The algorithm introduces a bias based on a cost matrix $C(i, j)$ in the training data and predicts the class with the minimum expected misclassification cost using the values in the cost matrix. Given an example, z and the probability $P(j|z)$ of each class j, the Bayes optimal prediction for z is the class i that minimizes the conditional risk in equation 14.

$$R(i|z) = \sum_j P(j|z)C(i, j) \tag{14}$$

6.2. Benchmarks

For benchmarking, we first use the method proposed by Marini et al. (2013) to obtain an initial result using our entire feature set. However, our primary benchmarks

Table 2: Cost matrix of pairwise class error weightings based on the degree of taxonomic relatedness of bird classes.

	a	b	c	d	e	f	g	h	i	j	k	l	m
a = Alexandrine Parakeet		0.50	0.50	1.00	0.75	0.75	1.00	0.75	1.00	1.00	0.75	0.75	1.00
b = Nanday Parakeet	0.50		0.50	1.00	0.75	0.75	1.00	0.75	1.00	1.00	0.75	0.75	1.00
c=Blue-crowned Parakeet	0.50	0.50		1.00	0.75	0.75	1.00	0.75	1.00	1.00	0.75	0.75	1.00
d=Common House Martin	1.00	1.00	1.00		1.00	1.00	0.75	1.00	1.00	1.00	1.00	1.00	0.75
e=Eastern Rosella	0.75	0.75	0.75	1.00		0.50	1.00	0.50	1.00	1.00	0.75	0.50	1.00
f=Budgerigar (yellow)	0.75	0.75	0.75	1.00	0.50		1.00	0.25	1.00	1.00	0.75	0.25	1.00
g =House Sparrow	1.00	1.00	1.00	0.75	1.00	1.00		1.00	1.00	1.00	1.00	1.00	0.75
h=Budgerigar (wild-type)	0.75	0.75	0.75	1.00	0.50	0.25	1.00		1.00	1.00	0.75	0.25	1.00
i=Common Wood Pigeon	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		1.00	1.00	1.00	1.00
j = Black Headed Gull	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00		1.00	1.00	1.00
k = Cockatiels	0.75	0.75	0.75	1.00	0.75	0.75	1.00	0.75	1.00	1.00		0.75	1.00
l=Budgerigar (blue)	0.75	0.75	0.75	1.00	0.50	0.25	1.00	0.25	1.00	1.00	0.75		1.00
m=Common Starling	1.00	1.00	1.00	0.75	1.00	1.00	0.75	1.00	1.00	1.00	1.00	1.00	

are single image classification using the VGG19 and MobileNet deep networks, which were used to compare our results after feature selection. A Linux server with three GeForce GTX TITAN X GPUs (12 GB memory each) was used to train all the networks. The networks were implemented using Python 3.5.3 and Keras 2.0.6 with Tensorflow backend. We used all of the images in the dataset with an 80/20 split for training and testing.

We used a transfer learning approach, utilising a version of the VGG-19 network pre-trained on the ImageNet dataset (Krizhevsky et al., 2012). The ImageNet dataset contains some bird classes, so we reasonably expect higher-level features in the pre-trained network to be applicable to ours. We replaced the input layer with our own (32x32x3) since many of our extracted silhouettes are small. We also replaced the fully connected (FC) layer with our own 13 output classes layer. We also introduced two dropouts, which were used to regularise the network to avoid over-fitting by randomly set a fraction (0.5) of FC layers (fc1 and fc2) units to zero at each update during training. The VGG19 model was trained using a stochastic gradient descent (SGD) optimizer, with a learning rate of 0.001 and momentum of 0.9. We used early stopping to interrupt training when the validation loss stops improving, by allowing a patience of 5 epochs. The pre-trained MobileNets model was also used: as with the VGG19 network, we removed the input layer (224x224x3) and replaced it with an input of shape (32x32x3). We replaced the FC layer with a new layer and introduced a dropout (0.5) to regularise the network to avoid over-fitting. MobileNet has the same setup as the VGG-19 network, except that we reduced the learning rate to 0.0001 and set the two additional hyper-parameters (width and resolution multipliers) to one.

7. Results

Table 3 shows the summary results of the four classifiers (NB, RF, RT and SVM) using our full feature set, alongside those obtained using Marini et al.'s features. The results show the overall correct classification rates, which are averaged across the 13 species for each classifier.

Table 3 suggests that the Random Forest classifier is superior, with the highest correct classification rate based on both our full feature sets and that of Marini et al.. Comparing ours with the method used by Marini et al. (2013) also aligns with our previous results in Atanbori et al. (2016).

Table 3: The overall correct classification rates (averaged across 13 classes) using **Marini et. al and our full feature set**, without feature selection for the four classifiers.

	Marini et. al	**Our feature Set**
Naive Bayes	49%	**69%**
Random Forest	77%	**83%**
Random Tree	62%	**66%**
SVM	57%	**74%**

We provide a confusion matrix for the RF classifier in Table 4.

7.1. Feature Selection

As mentioned, we have evaluated two methods of feature selection Hall (1999), and Breiman (2001), using all four classifiers. In this section, we describe how we determined the optimal subset and its composition. We used the following procedure to determine the optimal feature subset, for both methods:

- We used the feature selection method to rank the features.
- We iteratively removed the 10 lowest ranked features from the list.
- For each iteration, we evaluated the performance of each of the four classifiers, using the remaining features.
- This was repeated until only the 10 highest ranked features remained.
- We plotted a graph of correct classification rates and estimated the maxima for each classifier.

Figures 5 and 6 show the resulting graphs. Each classifier is shown as a separate curve. Peaks in the classifier-based method occur at 80 features for RF, 70 for RT and 100 for SVM and NB respectively. Likewise, peaks using the correlation-based method occurred at 80 features for RF classifier, and 70, 90 and 100 for RT, SVM and NB respectively.

We use RF for our analyses as overall it is the best performing classifier. Based on the RF, the mode of both feature selection techniques occurred at 70 features, which is 89% for the classifier-based method and 90% for the correlation-based method. Table 5 shows feature groups by type, before and after selection, for both methods. The classifier-based method selected 62 appearance features, from four feature groups and the 18 motion features from two groups. The correlation-based method selected 68 appearance features from four groups, and 12 motion features from one group. Wingbeat frequency was selected by both methods which suggest that they are significant for differentiating species. The highest ranked features for both classifier-based and correlation-based methods are the width of the FFT peaks (eight width peaks each). These relate directly to the periodicity of wing beats and therefore suggests that flight characteristics may be crucial for classifying species.

Birds are usually identified at close distance using their appearances (colour and shape). It is no surprise that most of the features selected were appearance features (62 for the classifier based method and 68 in the case of correlation-based). However, both methods selected only four shape features which strongly suggest the importance of the colour features compared with the shape.

Finally, our proposed feature reduction methods have shown improvement in performance with all classifiers apart from SVM. The results of these are therefore somewhat inconclusive, though we note that the improved performance with Random Forest is significant in the sense that this is consistently shown as the best classification method (with reduced features, full features, or Marini's features).

Table 6 shows the results of our full feature set obtained using correlation-based feature selection, which summarises the overall correct classification rate for each classifier. We have also provided a confusion matrix for RF (the best performing) in Table 7, to detail misclassification between species. Overall, the results of our full feature set with feature selection (90%) outperformed those without (83%), which asserts the efficacy of feature selection.

To help investigate misclassification between species, we considered Order and Family of the birds. We group species into:

1. Four Orders: **Psittaciformes**(Alexandrine Parakeet, Nanday Parakeet, Blue-crowned Parakeet, Budgerigar (wild-type), Budgerigar (yellow), Budgerigar (blue), Eastern Rosella and Cockatiel), **Charadriiformes**(Black-headed Gull), **Columbiformes**(Common Wood Pigeon) and **Passeriformes**(Common House Martin, Common Starling and House Sparrow).
2. Eight Families: **Psittacidae**(Alexandrine Parakeet, Nanday Parakeet and Blue-crowned Parakeet), **Psittaculidae**(Budgerigar (wild-type), Budgerigar (yellow), Budgerigar (blue) and Eastern Rosella), **Cacatuidae**(Cockatiel), **Laridae**(Black-

Table 4: The confusion matrix based on the **Random Forest** classifier without feature selection, using **the combined features** on the **thirteen classes dataset (ten bird species and another with three colour forms)**. %CC is the percentage correctly classified.

	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**	**% CC**	**Samples**
a = Alexandrine Parakeet	81.2%	6.4%	2.0%	0.5%	0.3%	1.3%	0.6%	0.7%	0.1%	1.8%	1.6%	3.2%	0.4%	**81%**	2611
b = Nanday Parakeet	9.3%	79.7%	2.4%	1.9%	0.3%	1.4%	0.2%	0.4%	0.3%	0.6%	1.7%	0.9%	0.7%	**80%**	2041
c=Blue-crowned Parakeet	8.7%	4.2%	82.2%	0.6%	0.4%	0.6%	0.3%	0.7%	0.1%	0.8%	0.2%	0.8%	0.2%	**82%**	1209
d=Common House Martin	0.1%	0.2%	0.0%	98.7%	0.0%	0.0%	0.3%	0.0%	0.0%	0.4%	0.1%	0.0%	0.2%	**99%**	6351
e=Eastern Rosella	3.8%	3.9%	2.6%	1.1%	66.3%	1.7%	1.0%	0.6%	0.0%	8.5%	2.8%	7.1%	0.6%	**66%**	1051
f=Budgerigar (yellow)	5.1%	2.8%	1.5%	1.2%	0.3%	80.9%	1.4%	1.2%	0.1%	2.6%	1.2%	1.2%	0.4%	**81%**	1202
g =House Sparrow	0.8%	0.1%	0.3%	19.5%	0.4%	0.9%	61.1%	0.5%	0.0%	5.6%	1.1%	1.3%	8.5%	**61%**	1952
h=Budgerigar (wild-type)	11.2%	4.7%	4.2%	2.9%	1.1%	6.7%	2.9%	47.5%	0.5%	9.6%	4.3%	3.5%	0.9%	**47%**	1097
i=Common Wood Pigeon	0.9%	1.0%	1.2%	3.3%	0.7%	0.7%	1.8%	0.8%	81.9%	2.4%	0.7%	0.3%	4.4%	**82%**	888
j = Black Headed Gull	0.1%	0.0%	0.0%	1.7%	0.1%	0.0%	0.1%	0.0%	0.0%	95.0%	1.5%	1.6%	0.0%	**95%**	7419
k = Cockatiels	1.4%	1.4%	0.1%	9.1%	0.4%	0.5%	1.4%	0.2%	0.1%	5.6%	74.8%	1.7%	3.3%	**75%**	1871
l=Budgerigar (blue)	2.6%	1.0%	0.3%	2.1%	0.9%	0.5%	1.3%	0.5%	0.0%	16.1%	2.0%	72.5%	0.3%	**72%**	2142
m=Common Starling	0.9%	0.8%	0.3%	3.9%	0.2%	1.0%	15.3%	0.2%	0.2%	1.2%	2.7%	0.4%	72.8%	**73%**	2180
Overall Correctly Classified														**83%**	**32014**

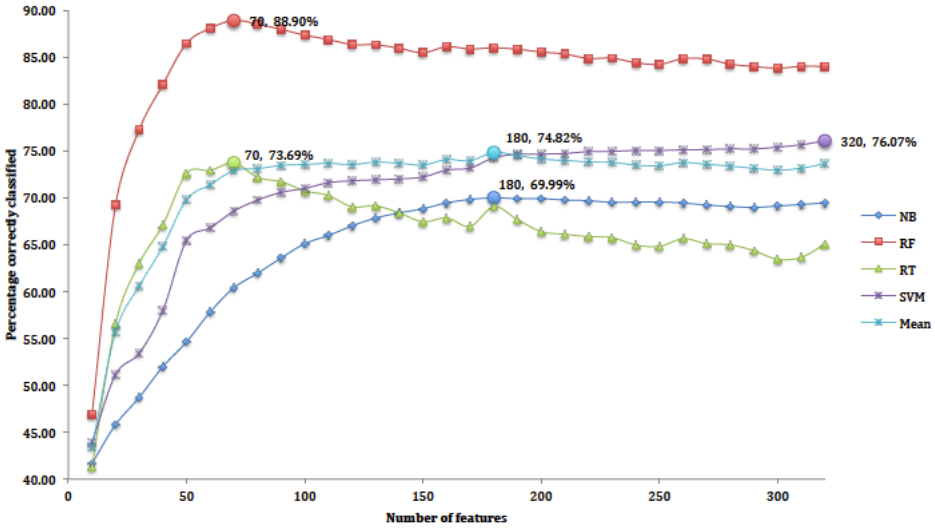

Figure 5: Classification rates vs. number of features using classifier-based selection. The maximum for each is marked with a solid circle, and labelled with the number of features.

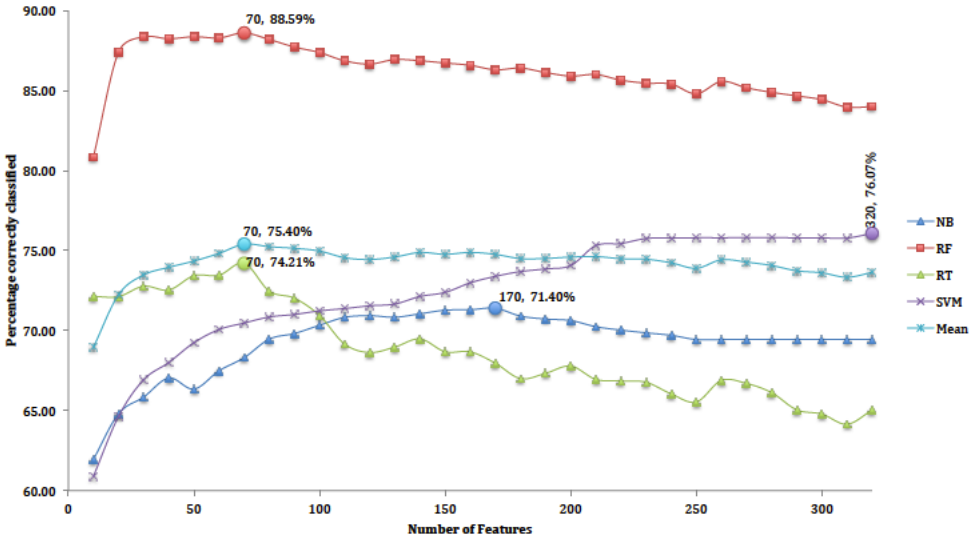

Figure 6: Classification rates vs. number of features for correlation-based selection. The maximum for each is marked with a solid circle, and labelled with the number of features.

Table 5: Number of features remaining in each group, before and after applying the classifier-based (CBfs) and correlation-based (CoBfs) Feature Selection, including the top features selected.

	Feature Group	# before FS	# after CBfs	# after CoBfs	**Top Selected Feature (CBfs)**	**Top Selected Feature (CoBfs)**
Appearance	Hue color features	37	22	18	σ of Hue	Mean of Hue
	Saturation colour features	35	15	15	Mean of Saturate	Mean of Saturate
	Value colour features	37	21	31	Entropy of value	Entropy of value
	Shape	17	4	4	Hu's First invariant	Hu's First invariant
	Gabor	20	0	0	N/A	N/A
	Grayscale	8	0	0	N/A	N/A
	LogPolar	15	0	0	N/A	N/A
Motion	FFT (Wingbeat)	27	16	12	First Peak of FFT (width)	First Peak of FFT (width)
	CSS	22	0	0	N/A	N/A
	CDF	10	0	0	N/A	N/A
	Turn	62	2	0	Turn ($\theta_{i=55}$)	N/A
	Vicinity	20	0	0	N/A	N/A
	Curvature	10	0	0	N/A	N/A
	Total Features	**320**	**70**	**70**		

headed Gull), **Columbidae**(Common Wood Pigeon), **Hirundinidae**(Common House Martin), **Sturnidae**(Common Starling) and **Passeridae**(House Sparrow)

We then aggregate the results from the confusion matrices of the thirteen classes to form a four-class order and eight class family confusion matrices (see results in supplementary materials). We grouped birds belonging to the same order together to create the order confusion matrices and likewise those belonging to the same family.

The full feature set without feature selection misclassified 9% Columbiformes as Passeriformes whiles this was only 4% with feature selection (see Table 1 in supplementary materials). The full feature set also misclassified Psittaciformes as Charadriiformes(6%) and Columbiformes as Psittaciformes (6%). These misclassifications reduced to 4% each with feature selection. The worse case misclassifications when considering family without feature selection is Passeridae versus Hirundinidae(19%) and Sturnidae versus Passeridae (15%). Again, with feature selection, these were reduced to 6% and 8% respectively.

7.2. Contribution of Motion Features to Classification Rates

Table 8 shows the results of using appearance versus full-features, with feature selection, which summarises the overall correct classification rate for each classifier. We have provided a confusion matrix for the Random Forest classifier (the best performing classifier) in Table 11, to detail misclassification between species when we use only appearance features. Results suggest that using both appearance and motion features yields the best classification rates.

We reconsider the challenge of species recognition using (poor quality) in-the-field data. Our method and those used for benchmarking are all challenged by similarities in appearance between particular species, casting the recognition problem as that of "fine-grained" classification. We propose that motion features can act as a weak classifier and assist with differentiation, and there is evidence to support this by comparing our combined motion and appearance feature set with that of Marini et al. in Table 3 (our features out-perform those of Marini et al. (2013) with all classifiers).

Table 6: Correct classification rates across all species, with and without correlation-based feature selection (FS) based on the full feature set.

	Without Feature Selection	**With Feature Selection**
Naive Bayes	69%	**71%**
Random Forest	83%	**90%**
Random Tree	66%	**77%**
SVM	**74%**	72%

Table 8: Correct classification rates across all species, with feature selection: Appearance versus the full feature set.

	Appearance Features	**Full features**
Naive Bayes	67%	**71%**
Random Forest	84%	**90%**
Random Tree	73%	**77%**
SVM	70%	**72%**

The contribution of motion features in resolving species may be considered limited with relatively short video clips: however, we note that this is likely to be the case in real in-the-field operation. Furthermore, we remark that the retained features in Table 5 include a number of motion features (wingbeat frequencies), which suggests that some of these features are significant. Indeed improvements in some species with a similar appearance are evident in Table 7, such as between Parakeet species, and between the Common Starling and House Sparrow. Overall, feature selection with the Random Forest classifier introduces a 7% improvement in classification.

We also considered the Order of birds and noted that the appearance feature set with feature selection misclassified more birds from different Orders compared with the full feature set (see Tables 7 and 8 in supplementary materials). For example, Columbiformes versus Psittaciformes (7%) and Columbiformes versus Passeriformes (9%) were both reduced to only 4% with the full feature set. The worse misclassification when considering family were Passeridae versus Hirundinidae(13%) and Passeridae vs Sturnidae (9%) with the appearance feature set (see Table 3 in supplementary materials). These were reduced to 6% and 7% respectively with the full feature set, which reconfirms the importance of motion features to aid classification of bird species.

Table 7: Confusion matrix based on the **Random Forest** classifier with **selected combined features**. Results using the correlation-based selection.

	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**	**% CC**	**Samples**
a = Alexandrine Parakeet	85.3%	5.2%	1.0%	0.5%	0.2%	1.0%	0.7%	0.8%	0.1%	1.1%	1.2%	2.7%	0.3%	**85%**	2611
b = Nanday Parakeet	7.6%	84.3%	1.8%	1.3%	0.7%	0.9%	0.1%	0.4%	0.3%	0.4%	0.9%	0.9%	0.5%	**84%**	2041
c=Blue-crowned Parakeet	5.2%	2.2%	88.1%	0.3%	0.5%	0.7%	0.3%	0.8%	0.1%	0.5%	0.2%	0.8%	0.2%	**88%**	1209
d=Common House Martin	0.1%	0.1%	0.0%	94.3%	0.0%	0.0%	3.9%	0.0%	0.0%	0.5%	0.1%	0.0%	1.0%	**94%**	6351
e=Eastern Rosella	2.1%	2.9%	1.9%	0.8%	75.5%	1.2%	0.9%	0.8%	0.1%	4.9%	1.6%	6.5%	0.8%	**76%**	1051
f=Budgerigar (yellow)	2.5%	2.0%	1.1%	0.6%	0.2%	87.5%	1.3%	0.7%	0.1%	1.6%	0.7%	1.4%	0.3%	**88%**	1202
g =House Sparrow	0.5%	0.2%	0.1%	5.8%	0.2%	0.8%	81.5%	0.4%	0.0%	1.9%	0.9%	1.0%	6.9%	**81%**	1952
h=Budgerigar (wild-type)	4.9%	1.8%	2.6%	0.5%	0.9%	1.8%	0.5%	79.0%	0.5%	0.4%	1.9%	4.1%	1.0%	**79%**	1097
i=Common Wood Pigeon	0.3%	0.5%	0.2%	1.5%	0.3%	0.7%	0.9%	0.8%	90.0%	2.0%	0.2%	0.7%	1.9%	**90%**	888
j = Black Headed Gull	0.1%	0.0%	0.0%	0.3%	0.1%	0.0%	0.1%	0.0%	0.0%	99.2%	0.1%	0.2%	0.0%	**99%**	7419
k = Cockatiels	0.5%	0.7%	0.1%	5.1%	0.5%	0.3%	0.9%	0.1%	0.0%	3.5%	83.6%	1.1%	3.6%	**84%**	1871
l=Budgerigar (blue)	1.8%	0.7%	0.2%	1.2%	0.7%	0.6%	1.3%	0.4%	0.0%	8.4%	1.3%	82.5%	0.9%	**83%**	2142
m=Common Starling	0.2%	0.1%	0.1%	2.8%	0.0%	0.3%	8.2%	0.2%	0.1%	1.2%	1.2%	0.5%	85.1%	**85%**	2180
Overall Correctly Classified														**90%**	32014

7.3. Results using the VGG19 and MolbileNet Image Classifiers

Tables 9 and 10 show the classification results obtained using the VGG-19 and MobileNet classifiers respectively. The overall correct classification rates for this classifiers were 84% and 80% respectively.

Comparison with the VGG-19 and MobileNet image classifiers are also noteworthy. VGG-19 outperformed MobileNet by 4%, which may be attributed to the fewer parameters of MobileNet since this deep learning model runs on devices with a limited resource (trading-off latency against accuracy). Both deep learning methods misclassified species with a similar shape. However, MobileNet misclassified more of these species than VGG-19. Another interesting observation is that MobileNet network misclassified more than 16% of Alexandrine Parakeet as Budgerigar (wild-type). Visualising the higher level filter of this model shows that it relies less on colour than VGG-19, which may explain some differences in performance.

Considering the overall classification accuracies using the random forest with feature selection (90% correct classification) outperformed both deep learning approaches. We analyse the results by examining the Order and Family of bird species to investigate this. The worse misclassification for both VGG-19 and MobileNet occurred with Psittaciformes versus Passeriformes (4%) and Charadriiformes versus Psittaciformes (5% for MobileNet and 2% for VGG), whiles our approach for these were 3% and 0% respectively. Our method also struggles with Columbiformes versus Passeriformes (4%) but VGG-19 was 3% and MobileNet, 2% for these Orders (see Tables 7, 9 and 10 in supplementary materials). Again ours misclassified Columbiformes versus Psittaciformes (4%) whiles these were respectively 1% for both VGG-19 and MobileNet. Both Deep Networks misclassified more Families especially Passeridae versus Hirundinidae (22% VGG-19 and 15% MobileNet) and Psittacidae versus Psittaculidae (12% VGG-19 and 21% MobileNet) whiles ours only misclassified 7% and 4% respectively (see Tables 2, 4 and 5 in supplementary materials). There were very few Families for which the deep learning methods were better than ours.

Whilst we do not claim that our approach will outperform all deep-learning methods, we do think that this evidences the potential of motion-based features to assist with the classification problem, especially when dealing with noisy data. We propose that future work could focus on exploring motion features over longer trajectories, and also look at network architectures which include temporal features.

Table 9: VGG-19 Confusion Matrix

	a	b	c	d	e	f	g	h	i	j	k	l	m	%CC	Samples
a=Alexandrine Parakeet	56.57%	5.44%	12.45%	0.54%	0.69%	1.26%	2.41%	13.10%	1.99%	0.04%	0.92%	4.21%	0.38%	57%	2611
b=Nanday Parakeet	4.12%	83.05%	4.41%	0.24%	0.24%	3.04%	0.05%	2.06%	2.40%	0.00%	0.29%	0.10%	0.00%	83%	2041
c=Blue-crowned Parakeet	1.57%	1.82%	80.40%	1.24%	3.56%	0.00%	2.65%	1.82%	1.24%	0.74%	1.65%	0.41%	2.89%	80%	1209
d=Common House Martin	0.00%	0.00%	0.02%	96.98%	0.00%	0.00%	1.24%	0.03%	0.00%	0.14%	0.14%	0.00%	1.45%	97%	6351
e=Eastern Rosella	2.09%	2.57%	1.90%	0.10%	70.22%	0.38%	4.57%	3.24%	0.29%	0.29%	2.09%	11.80%	0.48%	70%	1051
f=Budgerigar (yellow)	6.07%	0.42%	1.58%	0.00%	0.08%	82.28%	0.17%	7.90%	0.00%	0.08%	0.00%	1.41%	0.00%	82%	1202
g=House Sparrow	0.05%	0.00%	0.31%	0.87%	0.15%	0.15%	72.59%	0.61%	0.15%	0.10%	0.36%	2.51%	22.13%	73%	1952
h=Budgerigar (wild-type)	3.28%	2.83%	4.74%	0.64%	7.11%	33.27%	3.28%	30.54%	0.55%	0.64%	2.01%	10.85%	0.27%	31%	1097
i=Common Wood Pigeon	0.79%	0.23%	0.00%	0.45%	0.00%	0.00%	0.00%	0.00%	95.50%	0.34%	0.00%	0.34%	2.36%	95%	888
j=Black-headed Gull	0.34%	0.05%	0.00%	0.00%	0.31%	0.01%	0.08%	0.03%	0.11%	98.06%	0.23%	0.74%	0.04%	98%	7419
k=Cockatiel	1.87%	0.75%	0.32%	0.27%	2.83%	0.05%	0.91%	3.10%	1.12%	2.67%	74.56%	5.40%	6.15%	75%	1871
l=Budgerigar (blue)	0.93%	1.03%	0.33%	0.89%	1.07%	0.09%	1.59%	0.98%	0.42%	0.79%	0.28%	90.57%	1.03%	91%	2142
m=Common Starling	0.09%	0.00%	0.00%	0.87%	0.00%	0.09%	16.88%	0.18%	0.69%	0.50%	0.14%	0.37%	80.18%	80%	2180
Overall Correctly Classified														84%	

Table 10: MobileNet Confusion Matrix

	a	b	c	d	e	f	g	h	i	j	k	l	m	**%CC**	Samples
a=Alexandrine Parakeet	47.03%	4.52%	10.46%	0.23%	0.96%	3.64%	4.98%	16.47%	1.72%	0.38%	3.06%	6.24%	0.31%	**47%**	2611
b=Nanday Parakeet	2.25%	68.15%	5.14%	0.20%	4.31%	9.95%	0.10%	7.15%	1.42%	0.05%	0.98%	0.24%	0.05%	**68%**	2041
c=Blue-crowned Parakeet	2.07%	1.41%	82.55%	0.25%	4.14%	0.25%	2.81%	2.73%	0.33%	0.41%	2.32%	0.66%	0.08%	**83%**	1209
d=Common House Martin	0.00%	0.02%	0.03%	95.10%	0.00%	0.00%	1.97%	0.38%	0.03%	0.03%	1.45%	0.00%	0.99%	**95%**	6351
e=Eastern Rosella	0.38%	0.67%	1.33%	0.29%	62.13%	0.00%	5.42%	7.71%	0.00%	0.57%	3.71%	17.60%	0.19%	**62%**	1051
f=Budgerigar (yellow)	2.08%	0.33%	1.25%	0.00%	1.75%	86.69%	0.42%	4.66%	0.00%	0.42%	0.50%	1.91%	0.00%	**87%**	1202
g=House Sparrow	0.00%	0.00%	0.20%	2.51%	0.05%	0.46%	75.72%	2.10%	0.31%	0.20%	1.38%	1.69%	15.37%	**76%**	1952
h=Budgerigar (wild-type)	0.73%	0.64%	7.38%	0.18%	4.47%	34.46%	4.47%	34.82%	0.18%	0.46%	2.01%	9.75%	0.46%	**35%**	1097
i=Common Wood Pigeon	0.45%	0.45%	0.34%	0.45%	0.00%	0.00%	0.45%	0.11%	95.72%	1.01%	0.00%	0.00%	1.01%	**96%**	888
j=Black-headed Gull	0.04%	0.01%	0.01%	0.04%	0.12%	0.05%	0.61%	0.18%	0.03%	93.95%	1.78%	3.03%	0.15%	**94%**	7419
k=Cockatiel	0.21%	0.59%	0.00%	0.05%	1.87%	0.21%	3.05%	2.14%	1.07%	2.51%	79.32%	5.13%	3.85%	**79%**	1871
l=Budgerigar (blue)	2.15%	0.37%	0.75%	0.70%	1.26%	1.73%	2.24%	2.38%	1.54%	0.98%	0.42%	84.45%	1.03%	**84%**	2142
m=Common Starling	0.00%	0.14%	0.05%	1.06%	0.09%	0.09%	32.71%	0.46%	1.01%	0.18%	1.56%	0.46%	62.20%	**62%**	2180
Overall Correctly Classified														**80%**	

Table 11: Confusion matrix based on the **Random Forest** classifier with **selected appearance features only**. Results based on the correlation-based feature selection method.

	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**	**% CC**	**Samples**
a = Alexandrine Parakeet	84.3%	5.4%	1.2%	0.5%	0.3%	0.9%	0.7%	1.2%	0.2%	1.1%	1.3%	2.7%	0.3%	**84%**	2611
b = Nanday Parakeet	7.2%	83.5%	2.1%	1.1%	0.4%	1.3%	0.1%	0.7%	0.6%	0.6%	0.9%	0.8%	0.6%	**83%**	2041
c=Blue-crowned Parakeet	6.2%	2.2%	85.9%	0.2%	1.0%	0.8%	0.8%	1.0%	0.1%	0.2%	0.2%	1.1%	0.2%	**86%**	1209
d=Common House Martin	0.1%	0.2%	0.0%	90.0%	0.0%	0.0%	4.6%	0.0%	0.0%	4.4%	0.1%	0.1%	0.5%	**90%**	6351
e=Eastern Rosella	2.1%	2.0%	2.1%	0.4%	77.4%	0.8%	1.1%	1.0%	0.1%	4.2%	1.8%	6.8%	0.4%	**77%**	1051
f=Budgerigar (yellow)	2.4%	2.3%	0.9%	0.6%	0.2%	86.3%	1.5%	1.7%	0.1%	1.8%	0.6%	1.2%	0.3%	**86%**	1202
g =House Sparrow	0.2%	0.1%	0.1%	13.1%	0.3%	0.7%	68.4%	0.6%	0.0%	4.7%	1.1%	1.8%	8.8%	**68%**	1952
h=Budgerigar (wild-type)	7.1%	3.5%	2.8%	2.5%	1.1%	6.2%	3.6%	60.1%	0.5%	3.7%	3.4%	4.0%	1.5%	**60%**	1097
i=Common Wood Pigeon	0.8%	0.6%	0.7%	3.7%	0.7%	0.6%	0.8%	0.7%	81.4%	2.0%	1.9%	1.4%	4.8%	**81%**	888
j = Black Headed Gull	0.0%	0.0%	0.0%	1.6%	0.1%	0.0%	3.1%	0.6%	0.0%	90.4%	1.4%	2.7%	0.0%	**90%**	7419
k = Cockatiels	0.7%	0.5%	0.1%	5.1%	0.5%	0.5%	1.1%	0.4%	0.1%	2.5%	83.0%	1.2%	4.5%	**83%**	1871
l=Budgerigar (blue)	2.1%	0.7%	0.4%	1.1%	1.0%	0.6%	1.7%	1.0%	0.0%	9.1%	1.8%	80.0%	0.6%	**80%**	2142
m=Common Starling	0.4%	0.5%	0.2%	5.2%	0.1%	0.6%	9.3%	0.2%	0.2%	1.1%	2.4%	0.2%	79.6%	**80%**	2180
Overall Correctly Classified														**84%**	32014

8. Conclusion

We have presented our work on the automated classification of bird species in flight. Little comparable existing work addresses this problem domain. Those that do are mainly semi-automated, and use high-quality individual images, and so not suitable for field deployment. Ours is the first to adequately address this challenge as a robust fine-grained classification problem, and to consider combining motion and appearance features for classification in flight.

We defined a total set of 320 appearance and motion features, in conjunction with standard classifiers, on a thirteen classes dataset. The Random Forest classifier showed the best all-around performance (90% classification rate). We used feature selection to improve performance, which retained about 80 features and improving correct classification rates by around 7%. Comparison with state-of-the-art deep learning image classifiers (VGG, MobileNet), trained and tested on the same image frames of our video dataset, showed that our classification method compares favourably on our dataset.

Our future work focusses on further improving classifier performance, and expanding our dataset of species, through deployment in different contexts (such as migration). In particular, we consider that motion features will be most effective at a range were appearance features are not discernable. We wish to investigate the use of motion features with different temporal resolutions to improve species classification at this distances.

Acknowledgements

This work was supported by The National Parrot Sanctuary, Lincolnshire, UK, who have assisted with the collection of video data of several species used in this work.

References

Atanbori, J., Cowling, P., Murray, J., Colston, B., Eady, P., Hughes, D., Nixon, I., Dickinson, P., 2013. Analysis of bat wing beat frequency using fourier transform. In: Computer Analysis of Images and Patterns. Springer, pp. 370–377.

Atanbori, J., Duan, W., Murray, J., Appiah, K., Dickinson, P., September 2015. A computer vision approach to classification of birds in flight from video sequences. In: T. Amaral, S. Matthews, T. P. S. M., Fisher, R. (Eds.), Proceedings of the Machine Vision of Animals and their Behaviour (MVAB). BMVA Press, pp. 3.1–3.9.
URL https://dx.doi.org/10.5244/CW.29.MVAB.3

Atanbori, J., Duan, W., Murray, J., Appiah, K., Dickinson, P., 2016. Automatic classification of flying bird species using computer vision techniques. Pattern Recognition Letters 81, 53–62.

Berg, T., Belhumeur, P. N., 2013. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, pp. 955–962.

Berg, T., Liu, J., Lee, S. W., Alexander, M. L., Jacobs, D. W., Belhumeur, P. N., 2014. Birdsnap: Large-scale fine-grained visual categorization of birds. In: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. IEEE, pp. 2019–2026.

Betke, M., Hirsh, D. E., Makris, N. C., McCracken, G. F., Procopio, M., Hristov, N. I., Tang, S., Bagchi, A., Reichard, J. D., Horn, J. W., et al., 2008. Thermal imaging reveals significantly smaller brazilian free-tailed bat colonies than previously estimated. Journal of Mammalogy 89 (1), 18–24.

Beyan, Ç., 2015. Detection of unusual fish trajectories from underwater videos. PhD thesis, The University of Edinburgh.
URL https://www.era.lib.ed.ac.uk/bitstream/handle/1842/10561/Beyan2015.pdf

Beyan, C., Fisher, R. B., 2013. Detection of abnormal fish trajectories using a clustering based hierarchical classifier. In: British Machine Vision Conference (BMVC), Bristol, UK.

Branson, S., Van Horn, G., Belongie, S., Perona, P., 2014. Bird species categorization using pose normalized deep convolutional nets. arXiv preprint arXiv:1406.2952.

Breiman, L., 1996. Bagging predictors. Machine learning 24 (2), 123–140.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Bruderer, B., Peter, D., Boldt, A., Liechti, F., 2010. Wing-beat characteristics of birds recorded with tracking radar and cine camera. Ibis 152 (2), 272–291.

Chang, C.-C., Lin, C.-J., 2011. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2 (3), 27.

Cullinan, V. I., Matzner, S., Duberstein, C. A., 2015. Classification of birds and bats using flight tracks. Ecological Informatics 27, 55–63.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, pp. 248–255.

Domingos, P., 1999. Metacost: A general method for making classifiers cost-sensitive. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp. 155–164.

Du, J.-X., Wang, X.-F., Zhang, G.-J., 2007. Leaf shape based plant species recognition. Applied Mathematics and Computation 185 (2), 883–893.

Duan, K., Parikh, D., Crandall, D., Grauman, K., 2012. Discovering localized attributes for fine-grained recognition. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, pp. 3474–3481.

Duberstein, C., Virden, D., Matzner, S., Myers, J., Cullinan, V., Maxwell, A., 2012. Automated thermal image processing for detection and classification of birds and bats.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., Zisserman, A., 2012. The pascal visual object classes challenge 2012 (voc2012) results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Gavves, E., Fernando, B., Snoek, C. G., Smeulders, A. W., Tuytelaars, T., 2013. Fine-grained categorization by alignments. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1713–1720.

Gavves, E., Fernando, B., Snoek, C. G., Smeulders, A. W., Tuytelaars, T., 2015. Local alignments for fine-grained categorization. International Journal of Computer Vision 111 (2), 191–212.

Gonzalez, R. C., Woods, R. E., 2002. Digital image processing.

Gu, Q., Li, Z., Han, J., 2012. Generalized fisher score for feature selection. arXiv preprint arXiv:1202.3725.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. The Journal of Machine Learning Research 3, 1157–1182.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., 2009. The weka data mining software: an update. ACM SIGKDD explorations newsletter 11 (1), 10–18.

Hall, M. A., 1999. Correlation-based feature selection for machine learning. Ph.D. thesis, The University of Waikato.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Hristov, N. I., Betke, M., Theriault, D. E., Bagchi, A., Kunz, T. H., 2010. Seasonal variation in colony size of brazilian free-tailed bats at carlsbad cavern based on thermal imaging. Journal of Mammalogy 91 (1), 183–192.

Huang, C., Luo, B., Tang, L., Liu, Y., Ma, J., 2013. Topic model based bird breed classification and annotation. In: Communications, Circuits and Systems (ICCCAS), 2013 International Conference on. Vol. 2. IEEE, pp. 319–322.

Huang, Z. C., Chan, P. P. K., Ng, W. W. Y., Yeung, D. S., July 2010. Content-based image retrieval using color moment and gabor texture feature. In: 2010 International Conference on Machine Learning and Cybernetics. Vol. 2. pp. 719–724.

Jacob, M., Blu, T., Unser, M., 2001. An exact method for computing the area moments of wavelet and spline curves. IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (6), 633–642.

Krause, J., Jin, H., Yang, J., Fei-Fei, L., 2015. Fine-grained recognition without part annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5546–5555.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105.

Lazarevic, L., Harrison, D., Southee, D., Wade, M., Osmond, J., 2008. Wind farm and fauna interaction: detecting bird and bat wing beats through cyclic motion analysis. Int. Journal of Sustainable Engineering 1 (1), 60–68.

Lee, D., Redd, S., Schoenberger, R., Xu, X., Zhan, P. E., 2003. An automated fish species classification and migration monitoring system. In: Industrial Electronics Society, 2003. IECON'03. The 29th Annual Conference of the IEEE. Vol. 2. IEEE, pp. 1080–1085.

Lee, S., Park, Y.-T., dAuriol, B. J., et al., 2012. A novel feature selection method based on normalized mutual information. Applied Intelligence 37 (1), 100–120.

Lee, T. S., 1996. Image representation using 2d gabor wavelets. IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (10), 959–971.

Li, X., Hu, W., Hu, W., 2006. A coarse-to-fine strategy for vehicle motion trajectory clustering. In: Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. Vol. 1. IEEE, pp. 591–594.

Liwicki, M., Bunke, H., et al., 2006. Hmm-based on-line recognition of handwritten whiteboard notes. In: Tenth International Workshop on Frontiers in Handwriting Recognition.

Mai, F., Chang, C., Hung, Y., 2010. Affine-invariant shape matching and recognition under partial occlusion. In: Image Processing (ICIP), 2010 17th IEEE International Conference on. IEEE, pp. 4605–4608.

Marini, A., Facon, J., Koerich, A. L., 2013. Bird species classification based on color features. In: Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE, pp. 4336–4341.

Matzner, S., Cullinan, V. I., Duberstein, C. A., 2015. Two-dimensional thermal video analysis of offshore bird and bat flight. Ecological Informatics 30, 20–28.

Moore, J. H., White, B. C., 2007. Tuning relieff for genome-wide genetic analysis. In: Evolutionary computation, machine learning and data mining in bioinformatics. Springer, pp. 166–175.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27 (8), 1226–1238.

Pun, C.-M., Lee, M.-C., May 2003. Log-polar wavelet energy signatures for rotation and scale invariant texture classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (5), 590–603.

Robnik-Šikonja, M., Kononenko, I., 2003. Theoretical and empirical analysis of relieff and rrelieff. Machine learning 53 (1-2), 23–69.

Rodrigues, M. T., Pádua, F. L., Gomes, R. M., Soares, G. E., 2010. Automatic fish species classification based on robust feature extraction techniques and artificial immune systems. In: Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010 IEEE Fifth International Conference on. IEEE, pp. 1518–1525.

Rother, C., Kolmogorov, V., Blake, A., Aug. 2004. "grabcut": Interactive foreground extraction using iterated graph cuts. ACM Trans. Graph. 23 (3), 309–314.

Sergyan, S., Jan 2008. Color histogram features based image classification in content-based image retrieval systems. In: Applied Machine Intelligence and Informatics, 2008. SAMI 2008. 6th International Symposium on. pp. 221–224.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Spampinato, C., Giordano, D., Di Salvo, R., Chen-Burger, Y.-H. J., Fisher, R. B., Nadarajan, G., 2010. Automatic fish classification for underwater species behavior understanding. In: Proceedings of the first ACM international workshop on Analysis and retrieval of tracked events and motion in imagery streams. ACM, pp. 45–50.

Suzuki, S., Abe, K., 1985. Topological structural analysis of digitized binary images by border following 30 (1).

Suzuki, S., et al., 1985. Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing 30 (1), 32–46.

Tang, J., Alelyani, S., Liu, H., 2014. Feature selection for classification: A review. Data Classification: Algorithms and Applications, 37.

Toloşi, L., Lengauer, T., 2011. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics 27 (14), 1986–1994.

Wah, C., Branson, S., Perona, P., Belongie, S., 2011a. Multiclass recognition and part localization with humans in the loop. In: Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, pp. 2524–2531.

Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011b. The caltech-ucsd birds-200-2011 dataset.

Yu, L., Liu, H., 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: ICML. Vol. 3. pp. 856–863.

Zhang, Y., Wei, X.-s., Wu, J., Cai, J., Lu, J., Nguyen, V.-A., Do, M. N., 2015. Weakly supervised fine-grained image categorization. arXiv preprint arXiv:1504.04943.

Zivkovic, Z., van der Heijden, F., 2006. Efficient adaptive density estimation per image pixel for the task of background subtraction. Pattern recognition letters 27 (7), 773–780.