International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 27, No. 7 (2013) 1350020 (21 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0218001413500201

AUDIO-VISUAL RECOGNITION OF GOOSE FLOCKING BEHAVIOR

KIM ARILD STEEN

Department of Engineering
Aarhus University, Finlandsgade 22
8200 Aarhus N, Denmark
kas@iha.dk

OLE ROLAND THERKILDSEN

Department of Bioscience, Aarhus University
Grenåvej 14, 8410 Rønde, Denmark

OLE GREEN and HENRIK KARSTOFT

Department of Engineering
Aarhus University, Finlandsgade 22
8200 Aarhus N, Denmark

Received 7 March 2013
Accepted 30 July 2013
Published 25 September 2013

Every year, agriculture experience significant economic loss due to wild geese, rooks and other flocks of birds. A wide range of devices to detect and deter animals causing conflict is used to prevent this, although their effectiveness is often highly variable, due to habituation to disruptive or disturbing stimuli. Automated recognition of behaviors could form a critical component of a system capable of altering the disruptive stimulus to avoid habituation. This paper presents an audio-visual-based approach for recognition of goose flocking behavior. The vocal communication and movement of the flock is used for the audio-visual recognition, which is accomplished through classifier fusion of an acoustic and a video-based classifier. Acoustic behavior recognition is based on generalized perceptual features and support vector machines, and visual behavior recognition is based on optical flow estimation and a Bayesian Rule-Based scheme. Classifier fusion is implemented using the product rule on the soft-outputs from both classifiers. The algorithm has been used to recognize goose flocking behaviors (landing, foraging and flushing) and have improved the performance compared to using audio- or video-based classifiers alone. The improvement of using classifier fusion is most evident in the case of flushing and landing behavior recognition, where it was possible to combine the advantages of both the audio- and video-based classifier.

Keywords: Audio-visual; classifier fusion; intelligent agriculture; human–wildlife conflict; support vector machines; optical flow.

1. Introduction

Throughout the world, human–wildlife conflicts are increasing and today, wildlife management is an important part of modern agriculture.[28] Since damages caused by wildlife creates significant economic challenges, a wide range of devices to detect and deter animals causing conflict is used in wildlife damage management, although their effectiveness is often highly variable.[16] In most cases scaring devices are nonspecific, so they can be activated by any animal (through detection of motion and/or body heat[16]), not only when individuals of the target species enters the area. This increases the risk of habituation, which is often the major limitation on the use of scaring devices.[30] Although random or animal-activated scaring devices may reduce habituation and prolong the protection period over nonrandom devices,[30] no cost-effective concepts circumventing the problems of habituation have yet been developed. This paper presents an algorithm for automatic flocking behavior recognition. The recognition of flocking behavior enables us to assess the effect of scaring, by monitoring a subsequent change in behavior. Thereby our system may monitor potential habituation and, accordingly, change the disturbing or disruptive stimulus to circumvent this. A robust recognition of landing behavior is a crucial part of a scaring system, since geese are more alert during landing, and therefore highly susceptible to disruptive stimulus.

Both audio and video processing methods have been used for recognition of animal behavior.[19,23–25,29,37,38] Research within vocalization recognition has been highly influenced by methods conducted within human speech and speaker recognition. This includes feature extraction techniques, focused on cepstral features[22,33] and pattern recognition algorithms such as Hidden Markov Models (HMMs),[3,39] Gaussian Mixture Models (GMMs)[3] and Support Vector Machines (SVMs).[4,13,40] Computer vision techniques have been widely used in human/animal activity and behavior recognition.[9,19,41] In the case of a single or a few individuals, different approaches, such as the Kalman filter, Condensation algorithm or active shape models (ASM) have been used to track movement or model the posture in order to recognize behavior.[19,23,38] However, when considering crowd behavior, the motions of individuals within the crowd is the expression of a continuous flow that drives the crowd, and when a crowd is dense, usual tracking algorithms like Kalman filters or the Condensation algorithm generate large state space models, which are computationally expensive.[7]

Popular approaches in crowd motion estimation are background subtraction, temporal differencing and optical flow.[7,19,41] Optical flow, which is utilized in this paper, is an approximation of the motion in an image sequence, and has been used in behavior recognition, crowd motion simulations and event detection.[7,9,27]

The fusion of audio and video streams have been used in both speech recognition and human emotion recognition research.[32,42] In speech recognition, the fusion of spectral information in the audio channel and the tracking of lip movement have improved the performance of speech recognition, particularly in situations with low signal-to-noise ratio. In human emotion recognition, both the facial expression and the tonal information in the speech provide useful information about the state of mind of humans. In animal flocks, both the movement and the vocalizations, i.e. the communication within the flock, is often associated with certain behaviors. This makes a fusion of audio and video suitable for robust multi-modal recognition of animal flocking behavior.

There are different strategies for fusing audio and video streams. In human audio-visual speech recognition research, feature fusion and classifier fusion have been used to fuse the information from the two sources.[32] The most common method used in speech recognition is to perform feature fusion in a multi-stream hidden Markov model; however, we chose not to use HMM as a flock of geese produce vocalizations at the same time, resulting in a soundscape, where temporal information is not useful in the recognition. In contrast to feature fusion methods, the classifier fusion framework provides a mechanism for capturing the reliability of each modality, and thereby design the algorithm for robust recognition based on knowledge of the individual classifier performance. Here we chose classifier fusion as our fusion strategy based on the nature of the extracted features and the capability to design classifiers for each individual stream.

As such, this paper presents an algorithm for audio-visual-based recognition of flocking behavior. The vocal communication and movement of the flock is used for the audio-visual recognition, which is accomplished through classifier fusion of the acoustic- and video-based classifiers.

2. Materials and Methods

2.1. *Study species*

We chose the Russian/Baltic population of barnacle geese (*Branta leucopsis*) as our study subject. The dramatic increase in this population over the past few decades has led to serious conflict between agriculture and geese throughout the wintering range. In Denmark, the large flocks of barnacle geese, which occur along the west coast until late spring, are causing damage to both winter cereals and pastures. Moreover, barnacle geese, like other goose species, are vocal and therefore suitable for studying the relationship between vocalizations and behavior. Although various methods have been employed to scare barnacle geese off agricultural land, to date, no successful long-term, cost-effective scaring method has been found.

2.2. *Equipment*

A combination of a shielded shotgun microphone (Sennheiser MKE 400) and a machine vision camera (uEye UI-1245LE-C) with a field of view (FOV) of 45° connected to a laptop were used for recordings. A multiple-shielded audio extension cable was used to minimize loss in fidelity. The camera and laptop were placed in a box at the edge of the field, whereas the microphone was placed 10 m in front of the camera, closer to the geese. The system was powered by two 12 V 92 Ah deep cycling car batteries and data were stored on a 3 TB external hard drive. A detailed description can be found in Ref. 34.

2.3. *Data collection*

The vocalizations were recorded with a sample rate at 44.1 kHz. An uncompressed audio file (wave) was saved every five minutes during daylight hours. The synchronized audio and video recordings were stored on an external hard drive for later processing. In order to capture the movements of the geese, the video stream was recorded at a frame rate of 20 frames per second with a resolution of 640 × 480 pixels.[a] During the study period, there were two occurrences of barnacle geese, at two different dates, within the FOV of the camera. The recordings were categorized into the three behaviors of interest: landing, foraging and flushing. In Table 1, a description of the behaviors and the duration of the recordings are listed. The behaviors were observed as single events on both days, when the behaviors occurred. The behaviors were manually labeled[b] and the duration of certain behavior was based on subjective estimates, where the behavior of the majority of the flock was used to label the behavior. In the case of flushing behavior, the video material was very limited, as the geese quickly escaped the FOV of the camera. In Fig. 1, a single frame and a spectrogram of flushing behavior is shown. It is seen, that temporal information is not useful for acoustic recognition, as the spectrum does not change much over time (3.5 s). However, even though it is a still frame, it is clear that the geese are flying upwards, and the temporal information in their movement can be used for recognition of behavior. The recorded audio is plotted as a white waveform above the spectrogram.

Table 1. Description and duration of the recorded behaviors.

		Data	
Behavior	Definition	Audio	Video
No geese	No activity on the field	60 s	1200 frames
Landing	Multiple geese approach the field and land on the ground	241 s	4810 frames
Foraging	Multiple geese stay on the ground and feed	180 s	3600 frames
Flushing	Some geese take off, and the rest of the foraging flock follow, leaving the field empty of geese	9 s	180 frames

[a] Please contact Kim Arild Steen by email (kas@iha.dk) if you are interested in the data for further research.

[b] Ole Roland Therkildsen is a research biologist at Department of Bioscience, Aarhus University, and is an expert in goose ecology.

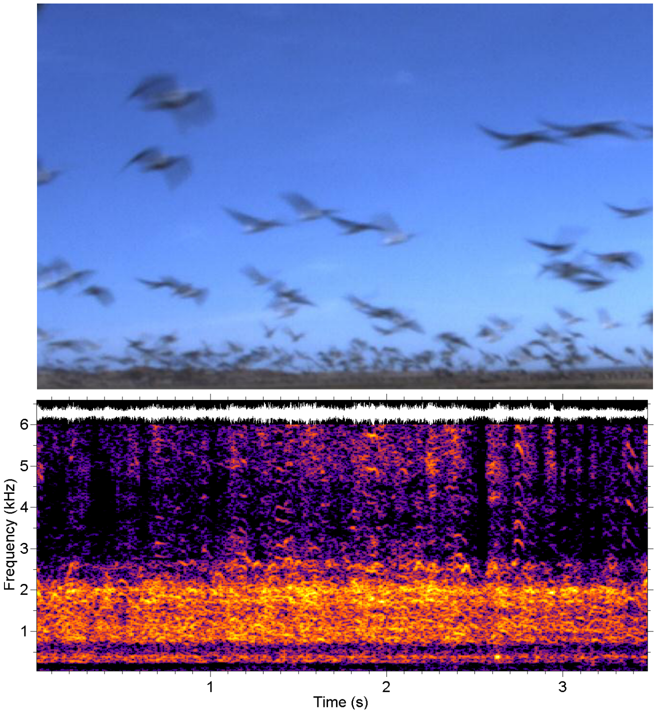


Fig. 1. A single frame and spectrogram of audio during flushing behavior. The behavior covers a short time span, and the acoustic channel has no useful temporal information, unlike the video channel.

3. Acoustic-Based Recognition

In Ref. 35, we presented an algorithm for acoustic-based recognition of goose behavior. The recognition was accomplished based on acoustic feature extraction and pattern recognition.

3.1. *Acoustic feature extraction*

The acoustic features were based on greenwood function cepstral coefficients (GFCC) which are similar to mel frequency cepstral coefficients (MFCC), which is often used in both human speech and animal vocalization recognition.[3,10,22] However, the frequency warping can be adjusted to the hearing capabilities of different species.[6]

Similarly to MFCC, the calculation of GFCC can be carried out using a scaled filter bank, consisting of a number of critical band filters with center frequencies adjusted to the specific scale.[10] The number of filters in the filter bank depend on the application, and various implementations of MFCC feature extraction have been used in speech recognition tasks.[15] The bandwidth of these applications differ, and as barnacle goose vocalizations contain most of their spectral information in the 500–6000 Hz band,[11] it is comparable to the bandwidth used in Ref. 8, where 20 filters were used. Therefore, 20 filters were used in the feature extraction of goose vocalizations. The calculation of GFCC was carried out as a short time analysis. We used 46 ms (2048 samples), windowed with a hamming window with the same length. Spectral vectors (S_k) represent the log-energy of each critical band, and a cosine transform converts the spectral vectors into cepstral vectors, according to the formula

$$c_n = \sum_{k=0}^{K-1} S_k \cos\left(n\left(k-\frac{1}{2}\right)\frac{\pi}{K}\right), \quad n = 0, \ldots, K-1. \tag{1}$$

Here c_n is the nth cepstral coefficients and S_k is the spectral log-energy of the kth band.

3.2. *Support vector machines*

In Ref. 35, we used SVM with a radial basis kernel function to classify behaviors. The SVM was chosen over the more popular HMM, as temporal information in the goose vocalizations is lost as multiple geese vocalize at the same time. Recently, SVM models have proven succesful in bird species recognition research,[13,40] and other research working with real-world classification tasks.[26] In addition, SVM models are able to handle nonlinear classification tasks, which is true for most real-world data.

SVM produce a *crisp* value, meaning that the output is either one class or another. Given the similarities in soundscapes for landing and flushing behavior, this may produce erroneous recognition results as data may often be close to the decision boundary. Soft-outputs/probability measures make it possible to detect these situations, and this information may be employed in the fusion of the classifier outputs. Here, we use LibSVM, which makes it possible to obtain probability estimates from the SVM classification.[5] This is accomplished by estimating the class probabilities, for k classes given any data $\mathbf{x}$

$$p_i = P(y = i|\mathbf{x}), \quad i = 1, \ldots, k. \tag{2}$$

The algorithm for accomplishing this is described in detail in Refs. 5 and 31. Based on the probability estimates for each class given the data $\mathbf{x}$, the soft-output recognition scheme for audio stream is found by the directional graph[13,31,35] shown in Fig. 2. The resulting output is not a *crisp*, but a measure of the probability of the three behaviors and the state of no geese.

The data $\mathbf{x}$ is a feature vector extracted from a short time audio sequence (46 ms), which is a very short time window for behavior recognition, since behavior should be estimated over a longer period of time. Therefore, we concatenate seven short time audio sequences into a matrix, with dimensions 7 × 6 (the dimensionality of features were reduced to six based on feature selection in Ref. 35), based on 0.2 s of audio data.

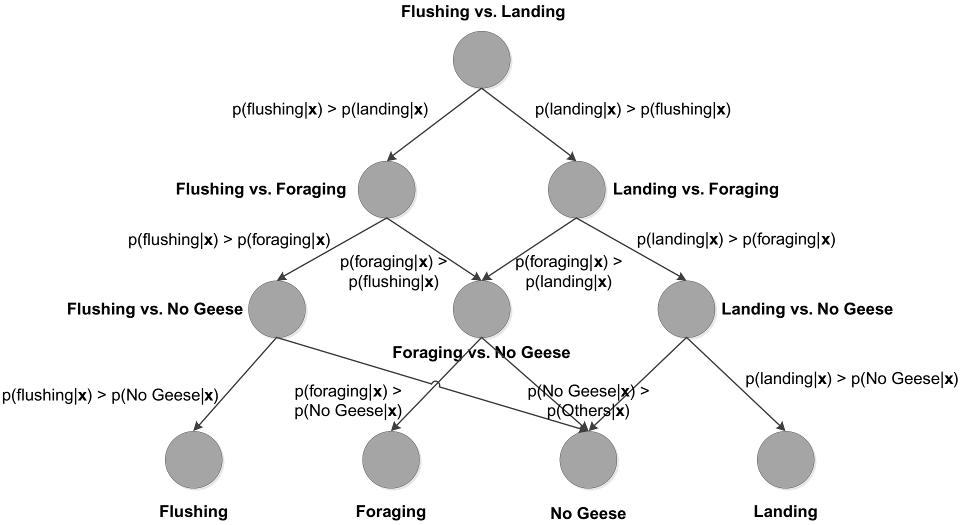

Flushing vs. Landing
p(flushing|**x**) > p(landing|**x**)
p(landing|**x**) > p(flushing|**x**)
Flushing vs. Foraging
Landing vs. Foraging
p(flushing|**x**) > p(foraging|**x**)
p(foraging|**x**) > p(flushing|**x**)
p(foraging|**x**) > p(landing|**x**)
p(landing|**x**) > p(foraging|**x**)
Flushing vs. No Geese
Foraging vs. No Geese
Landing vs. No Geese
p(flushing|**x**) > p(No Geese|**x**)
p(foraging|**x**) > p(No Geese|**x**)
p(No Geese|**x**) > p(Others|**x**)
p(landing|**x**) > p(No Geese|**x**)
Flushing
Foraging
No Geese
Landing

Fig. 2. Given a data point $\mathbf{x}$ the probability of each behavior and the state of no geese is found through a directional graph.

The final soft output is found by the mean of five of the seven outputs, as the smallest and largest values are removed to avoid outliers.

4. Video-Based Recognition

In Ref. 36, we presented an algorithm for recognition of flocking behavior using computer vision methods. The algorithm was based on optical flow estimation and a Rule-Based Bayesian scheme.

4.1. *Optical flow*

Optical flow is an approximation of the motion in an image sequence, given by velocity estimates. In crowd or flocking behavior, the motions of individuals within the flock is the expression of a continuous flow that drives the crowd, which can be estimated by the optical flow algorithm. The underlying assumption when computing optical flow is that pixel intensities are translated from one frame to the next by

$$I(\mathbf{x}, k) = I(\mathbf{x} + \mathbf{u}, k + 1), \tag{3}$$

where $I(\mathbf{x}, k)$ is the intensity of the pixel, located in the coordinate $\mathbf{x} = (i, j)^T$ at time k. The vector $\mathbf{u} = (u_x, u_y)^T$ is the 2D velocity vector. The above assumption rarely holds, since pixel brightness may change due to object rotations and secondary illumination. However, the assumption works well in practice.[14] The translation in Eq. (3) can be expressed as the *gradient constraint equation*[14] (4)

$$\nabla I(\mathbf{x}, k) \cdot \mathbf{u} + I_k(\mathbf{x}, k) = 0, \tag{4}$$

where $\nabla I \equiv (\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}) \equiv (I_i, I_j)$ and I_k denotes spatial and temporal partial derivate of the image. The above equation holds two unknowns, and further constraints are therefore necessary to estimate the velocity $\mathbf{u} = (u_x, u_y)^T$. In Ref. 18, Horn and Schunck combined the gradient constraint (4) with a global smoothness term to constrain the estimated velocity, minimizing

$$E(\mathbf{u}) = \int (\nabla I \cdot \mathbf{u} + I_k)^2 + \lambda(\|\nabla u_x\|^2 + \|\nabla u_y\|^2) dxdy, \tag{5}$$

where λ is a regularizing smoothness term. The estimated velocity is found by an iterative procedure, finding the minimum of (5). Here we used $\lambda = 0.5$ and at most 100 iterations per frame, as in Ref. 1, where the optical flow algorithm was used on both real and synthetic data.

4.2. *Behavior recognition*

The implementation of the Rule-Based scheme is presented in Fig. 3, where a flowchart of the recognition scheme is presented. First, the presence of geese is evaluated. This is based on the magnitude of movement, given by $v_k(i,j) = \|\mathbf{u}_k(i,j)\|$, present in the current frame. The magnitude of movement is an image constructed from the velocity vector estimates at each pixel location (i,j). If no geese are present, the magnitude has a low value and the presence of geese can be determined by a threshold on $\mathbf{v}$, denoted by Th in the figure, and given by the region R_k, which are the pixels in the image at time k above the threshold Th (6). If geese are present, the three behaviors are estimated based on probability measures derived from the estimated optical flow, as this gives a measure of the direction and magnitude of the majority of the velocity vectors.

$$R_k = \{(i,j) | v_k(i,j) > \text{Th}\}. \tag{6}$$

Since a change of goose behavior takes place over several frames, it is not suitable to classify behavior based on one frame alone. To incorporate information of

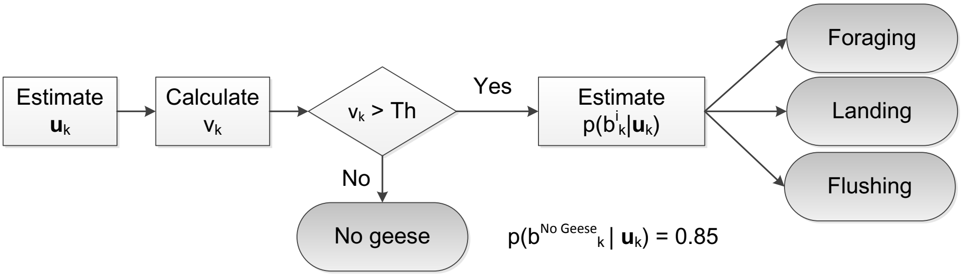


Fig. 3. Flow of the Rule-Based Bayesian scheme. First, the optical flow algorithm is used to estimate the velocity vectors, and the presence of geese is determined based on the magnitude of movement. The probability of the state of no geese is set to a fixed value of 0.85, as the presence is based on a threshold, which gives a crisp output, and the fusion requires a soft output. If geese are present, the probability of each behavior is estimated.

behavior from previous frames, we propose a Bayesian scheme where probability estimates from previous frames act as prior information. This is implemented using Bayes rule (7)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \tag{7}$$

where $P(B|A)$ and $P(A)$ are the likelihood and prior, respectively. The denominator $P(B)$, sometimes called evidence,[2] ensures that the posterior probability $P(A|B)$ is a valid probability measure. In our application, B denotes data, which is given by $\mathbf{u}_k$ and A denotes behavior, which is given by $b_k^i, i = \{\textit{No Geese}, \textit{Foraging}, \textit{Flushing}, \textit{Landing}\}$. At each frame, the posterior probability of a behavior given the data, is updated based on prior information of this behavior and the current probability of data given a specific behavior, which is defined as the following:

The probability $p(\mathbf{u}_k|b_k^{\text{foraging}})$ at time k is based on the movement at ground level G_k (10) compared to the general movement within the image.

$$G_k = \{(i,j)|j \leq j_0\}. \tag{8}$$

Only velocity vectors resulting in a large magnitude of movement are used to estimate the probability, as small magnitudes could arise from random noise in the image leading to a noisy probability estimate. The probability is found by (9)

$$p(\mathbf{u}_k|b_k^{\text{foraging}}) = \frac{\sharp(R_k \cap G_k)}{\sharp R_k}, \tag{9}$$

where the notation $\sharp$ denotes the counting operator.

The ground level is defined as the lower 40 pixel rows in the 640 × 480 pixels image, given by j_0. The part of the image, which is not ground level, is denoted sky level S_k.

$$S_k = \{(i,j)|j > j_0\}. \tag{10}$$

In the scope of this paper, the parameter j_0 is a fixed value. Methods to automatically find this value will be addressed in the discussion section.

The probability of data containing landing and flushing behavior is based on the direction of the velocity vectors in S_k, which consist of upward and downward directed velocity vectors (11).

$$S_k = S_k^{\uparrow} \cup S_k^{\downarrow}. \tag{11}$$

The directionality is found by evaluating the sign of $u_{y,k}$, which gives the following

$$S_k^{\uparrow} = S_k \cap \{(i,j)|u_{y,k}(i,j) \geq 0\}, \tag{12}$$

$$S_k^{\downarrow} = S_k \cap \{(i,j)|u_{y,k}(i,j) < 0\}. \tag{13}$$

The probabilities $p(\mathbf{u}_k|b_k^{\text{flushing}})$ and $p(\mathbf{u}_k|b_k^{\text{landing}})$ are estimated in the same manner as with foraging behavior:

$$p(\mathbf{u}_k|b_k^{\text{flushing}}) = \frac{\sharp(R_k \cap S_k^{\uparrow})}{\sharp R_k}, \tag{14}$$

$$p(\mathbf{u}_k|b_k^{\text{landing}}) = \frac{\sharp(R_k \cap S_k^{\downarrow})}{\sharp R_k}. \tag{15}$$

Using Bayes rule, with the probability estimates from the previous frame as prior, the posterior probability ($p(b_k^i|\mathbf{u}_k)$) of a single behavior, is based on estimates from previous frames (16)

$$p(b_k^i|\mathbf{u}_k) = \frac{p(\mathbf{u}_k|b_k^i)p(b_{k-1}^i)}{\sum_j p(\mathbf{u}_k|b_k^j)p(b_{k-1}^j)}. \tag{16}$$

This provides a soft-output from the video stream, which can be used in the classifier fusion described in Sec. 5.1.

5. Audio-Visual Behavior Recognition

The recognition of behavior is based on the methods described in the two previous sections, and a flow describing the procedure of the behavior recognition, is shown in Fig. 4. The vocalizations are recorded and divided into short time sequences and probability estimates for each behavior is found via SVM. These estimates are fused with the probability estimates from the video stream. The probability estimates from the video stream are estimated at frame level based on optical flow measures and a Bayesian Rule-Based scheme.

The two streams are synchronized at the classifier fusion level, based on the frame count of the video stream. For every four frames (0.2 s), the estimates from each stream are fused and a single behavior is classified based on the fusion described in Sec. 5.1.

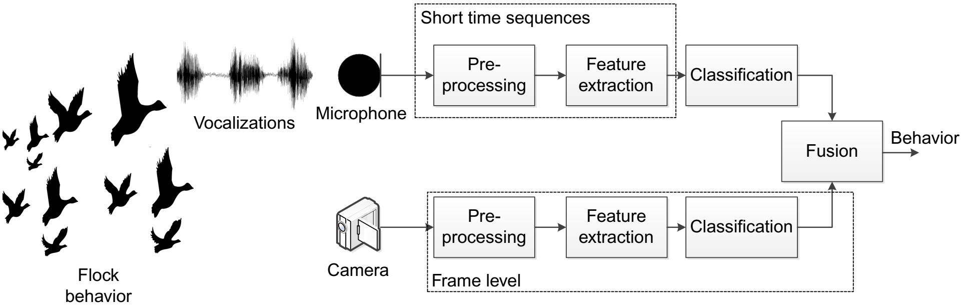


Fig. 4. The flow of audio-visual behavior recognition.

5.1. *Classifier fusion*

The audio-visual fusion is accomplished through classifier fusion, where the outputs (crisp or soft) from multiple classifiers are combined. The result is a single output, which is the combined decision of the classifiers. Previously, classifier fusion have been used for various tasks, including audio-visual speech recognition and human emotion recognition. Classifier fusion can be performed with different classifiers on the same data set, using different training data with the same type of classifiers or different data sets, as in audio-visual fusion.[20,32,42] The outputs from the classifiers may be used as features for a combining classifier or a rule-based fusion. Here we used a rule-based fusion to implement the fusion, and the result of using the sum, product and mean rule is presented in this paper. The rule-based approach is fast and easy to implement, and it is easy to introduce stream specific scaling of the soft-outputs, based on the performance of the individual classifiers. The task of the classifier fusion is as follows. Given R classifiers (where $R = 2$ in this paper), the pattern Z can be assigned to m possible classes $\{\omega_1, \dots, \omega_m\}$ by

$$\begin{aligned} &\text{assign } Z \to \omega_j \quad \text{if} \\ &P(\omega_j|\mathbf{x}_1, \dots, \mathbf{x}_R) = \max\nolimits_k P(\omega_k|\mathbf{x}_1, \dots, \mathbf{x}_R), \end{aligned} \tag{17}$$

where $P(\omega_j|\mathbf{x}_1, \dots, \mathbf{x}_R)$ denotes the probability of class ω_j given the data, $\mathbf{x}$, from R different classifiers. The rules implemented and tested in this paper are all presented in detail in Ref. 20.

Product rule. Under the assumption of statistical independence, the posterior probability of m different classes, given the outputs from R classifiers, can be expressed as a product of the individual conditional probabilities.[20] The product rule assigns the pattern Z to the class ω_j if the product of the posterior probabilities from the individual classifiers for class ω_j equals the maximum posterior probability. Written in mathematical terms in (18).

$$\begin{aligned} &\text{assign } Z \to \omega_j \quad \text{if} \\ &P^{-(R-1)}(\omega_j) \prod_{i=1}^{R} p(\omega_j|\mathbf{x}_i) = \max_{k=1}^{m} P^{-(R-1)}(\omega_k) \prod_{i=1}^{R} p(\omega_k|\mathbf{x}_i). \end{aligned} \tag{18}$$

Sum rule. Under the assumption that the *a posteriori*probabilities computed by the different classifiers will not deviate much from the prior probabilities, the above principles can be repeated using a summation of the posterior probabilities. It is a rather strong assumption, but it may be satisfied when the available data is highly ambiguous due to high levels of noise.[20] The sum rule is given by (19)

$$\begin{aligned} &\text{assign } Z \to \omega_j \quad \text{if} \\ &(1-R)P(\omega_j) \sum_{i=1}^{R} p(\omega_j|\mathbf{x}_i) = \max_{k=1}^{m} \left[(1-R)P(\omega_k) + \sum_{i=1}^{R} p(\omega_k|\mathbf{x}_i) \right]. \end{aligned} \tag{19}$$

Mean rule. Under equal prior assumptions, the sum rule (19) can be used to compute the average *a posteriori* probability for each class over all the classifier outputs. The mean rule assigns a pattern to that class for which the average *a posteriori* probability is maximum (20).

$$\text{assign } Z \to \omega_j \quad \text{if}$$
$$\frac{1}{R}\sum_{i=1}^{R} p(\omega_j|\mathbf{x}_i) = \max_{k=1}^{m} \frac{1}{R}\sum_{i=1}^{R} p(\omega_k|\mathbf{x}_i). \tag{20}$$

When using soft outputs it is possible to scale the output based on the performance of the individual classifier. In the literature, this is known as reliability ratio/measures, which is often used in human audio-visual speech recognition, where low signal-to-noise ratio in the audio stream would degrade performance.[12,32]

Like classifier fusion, there exist various strategies and methods for estimating the reliability of the different classifiers. The ratio can be estimated based on performance during training of the individual classifiers, and set as static weights on the output of the specific classifier. Another approach is to estimate the reliability via probabilistic measures in a Bayes network, based on features, signal quality, etc.,[21] and other more or less adaptive strategies. Here, the reliability ratio is estimated based on how confident the specific classifier is on its estimate of the behavior. Let $\rho(\omega_*)$ and $\rho(\omega_{**})$ denote the best and the second best probability estimate from a single classifier, the certainty λ is found by (21) (notation is inspired by Ref. 12).

$$\lambda = \rho(\omega_*) - \rho(\omega_{**}). \tag{21}$$

The weights (λ_{audio} and λ_{video}) found by (21) is multiplied to the output of the specific classifier before the fusion rule is applied. The framework of the classifier fusion enables prior information of behavior. Like the video-based classifier, the prior information is based on knowledge from previous classifications. Here, we denote this with superscript $t-1$:

$$\begin{aligned}&\text{assign } Z \to \omega_j \quad \text{if}\\ &\quad \lambda_{1,\ldots,R} \cdot P^{t-1}(\omega_j) \cdot P^{t}(\omega_j|\mathbf{x}_1,\ldots,\mathbf{x}_R)\\ &\quad = \max_k(\lambda_{1,\ldots,R} \cdot P^{t-1}(\omega_j) \cdot P^{t}(\omega_k|\mathbf{x}_1,\ldots,\mathbf{x}_R)).\end{aligned} \tag{22}$$

6. Results

The presented methods have been applied to the data described in Sec. 2.3. The performance have been tested via a 5-fold cross validation technique. We choose to omit the randomization in the cross validation, as the video-based classifier depends on frame-by-frame comparisons, and randomization of data would violate this. The results shown in Tables 2 and 3 are the mean performance ± standard deviation from the five folds.

Table 2. Normalized confusion matrix showing the performance of audio, video and classifier fusion, Sum, Product and Mean, (mean ± s.d.). NG = No Geese, FL = Flushing, L = Landing and FO = Foraging.

		Predicted			
Observed	C	NG	FL	L	FO
NG	A	**0.95 ± 0.05**	0 ± 0	0.02 ± 0.02	0.02 ± 0.04
	V	**0.88 ± 0.01**	0.11 ± 0.01	0.01 ± 0	0 ± 0
	S	**1 ± 0**	0 ± 0	0 ± 0	0 ± 0
	P	**1 ± 0**	0 ± 0	0 ± 0	0 ± 0
	M	**1 ± 0**	0 ± 0	0 ± 0	0 ± 0
FL	A	0 ± 0	**0.71 ± 0.22**	0.25 ± 0.21	0.03 ± 0.08
	V	0 ± 0	**1 ± 0**	0 ± 0	0 ± 0
	S	0 ± 0	**0.92 ± 0.17**	0.08 ± 0.17	0 ± 0
	P	0 ± 0	**0.9 ± 0.17**	0.08 ± 0.17	0.02 ± 0.04
	M	0 ± 0	**0.92 ± 0.17**	0.08 ± 0.17	0 ± 0
L	A	0.06 ± 0.08	0.03 ± 0.08	**0.88 ± 0.07**	0.02 ± 0.03
	V	0.01 ± 0.01	0.33 ± 0.32	**0.65 ± 0.33**	0.01 ± 0.01
	S	0.01 ± 0.01	0 ± 0	**0.99 ± 0.01**	0 ± 0
	P	0.01 ± 0.01	0 ± 0	**0.99 ± 0.01**	0 ± 0
	M	0.01 ± 0.01	0 ± 0	**0.99 ± 0.01**	0 ± 0
FO	A	0.01 ± 0.01	0 ± 0	0.07 ± 0.08	**0.92 ± 0.08**
	V	0 ± 0	0.01 ± 0.02	0.08 ± 0.17	**0.91 ± 0.17**
	S	0 ± 0	0 ± 0	0 ± 0	**1 ± 0**
	P	0 ± 0	0 ± 0	0 ± 0	**1 ± 0**
	M	0 ± 0	0 ± 0	0 ± 0	**1 ± 0**

Table 3. Comparison of performance using audio, video or classifier fusion, Sum, Product and Mean (mean ± s.d.). NG = No Geese, FL = Flushing, L = Landing and FO = Foraging.

		Performance		
Behavior	C	Accuracy[a]	Specificity[b]	Sensitivity[c]
NG	A	0.97 ± 0.03	0.97 ± 0.03	0.95 ± 0.05
	V	0.96 ± 0.03	0.99 ± 0.01	0.88 ± 0.01
	S	0.99 ± 0.01	0.99 ± 0.01	1 ± 0
	P	0.99 ± 0.01	0.99 ± 0.01	1 ± 0
	M	0.99 ± 0.01	0.99 ± 0.01	1 ± 0
FL	A	0.91 ± 0.05	0.98 ± 0.02	0.71 ± 0.22
	V	0.88 ± 0.08	0.84 ± 0.11	1 ± 0
	S	0.98 ± 0.04	1 ± 0	0.92 ± 0.17
	P	0.98 ± 0.04	1 ± 0	0.9 ± 0.17
	M	0.98 ± 0.04	1 ± 0	0.92 ± 0.17
L	A	0.88 ± 0.07	0.88 ± 0.1	0.88 ± 0.07
	V	0.89 ± 0.08	0.97 ± 0.06	0.65 ± 0.33
	S	0.97 ± 0.04	0.97 ± 0.06	0.99 ± 0.01
	P	0.97 ± 0.04	0.97 ± 0.06	0.99 ± 0.01
	M	0.97 ± 0.04	0.97 ± 0.06	0.99 ± 0.01

Table 3. (*Continued*)

		Performance		
Behavior	C	Accuracy[a]	Specificity[b]	Sensitivity[c]
FO	A	0.96 ± 0.03	0.97 ± 0.03	0.92 ± 0.08
	V	0.97 ± 0.04	0.99 ± 0.01	0.91 ± 0.17
	S	1 ± 0	1 ± 0	1 ± 0
	P	0.99 ± 0.01	0.99 ± 0.01	1 ± 0
	M	1 ± 0	1 ± 0	1 ± 0

[a] Ratio of correct predictions (both positive and negative) that were correct.
[b] Ratio of correct negative predictions (the ability to reject).
[c] Ratio of correct positive predictions.

The classification results are presented in a normalized confusion matrix (Table 2), which gives the ratio of correct positive predictions (as bold numbers) and correct negative predictions, where the classifier rejects a behavior correctly. Positive predictions or negative predictions, which are incorrect, are also given in the table. The performance of the models are given in Table 3, by three measures: accuracy, precision and sensitivity.

In Table 2 it is seen that the three fusion methods performs almost similar, and they all improve the overall performance in behavior classification. In the case of flushing behavior, the audio-based classifier, which relies on a trained model,[c] does not perform well compared to the video-based classifier. The fusion improves this, as both mean and standard deviation show better performance.

The fusion slightly degrades the performance of the video-based classifier w.r.t. sensitivity, and this will be addressed in the discussion section. In Table 3 it is seen that even though the sensitivity is degraded compared to the video-based classifier, both the accuracy and specificity has been improved.

In the case of landing behavior, the roles have changed, and the fusion improves the video-based classifier, which is more confused when detecting landing behavior, due to the nature of this specific behavior. In Table 2 it is seen that both the mean and standard deviation are improved, and Table 3 also show improved accuracy and specificity, when detecting landing behavior.

7. Discussion

Based on the results it is seen that fusion of audio and video results in an improvement of the recognition of goose flocking behaviors.

The video-based recognition performs worse than the audio based in the case of landing behavior. This is due to the nature of landing behavior. Not all geese land at the same time and some geese might take off again to find a better location. This affects the probability of landing, as optical flow estimation finds both downward

[c] Very little flushing data was available.

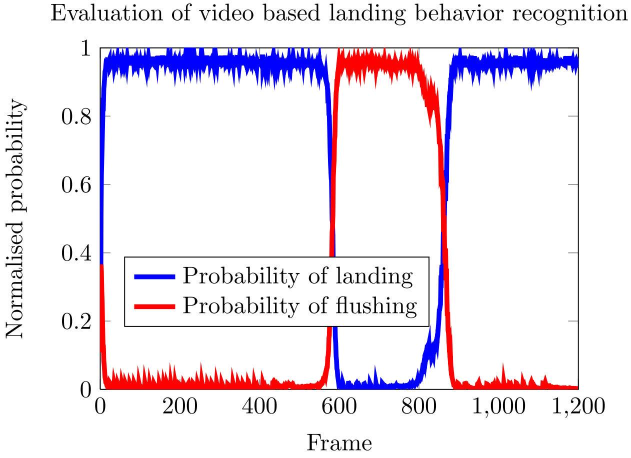


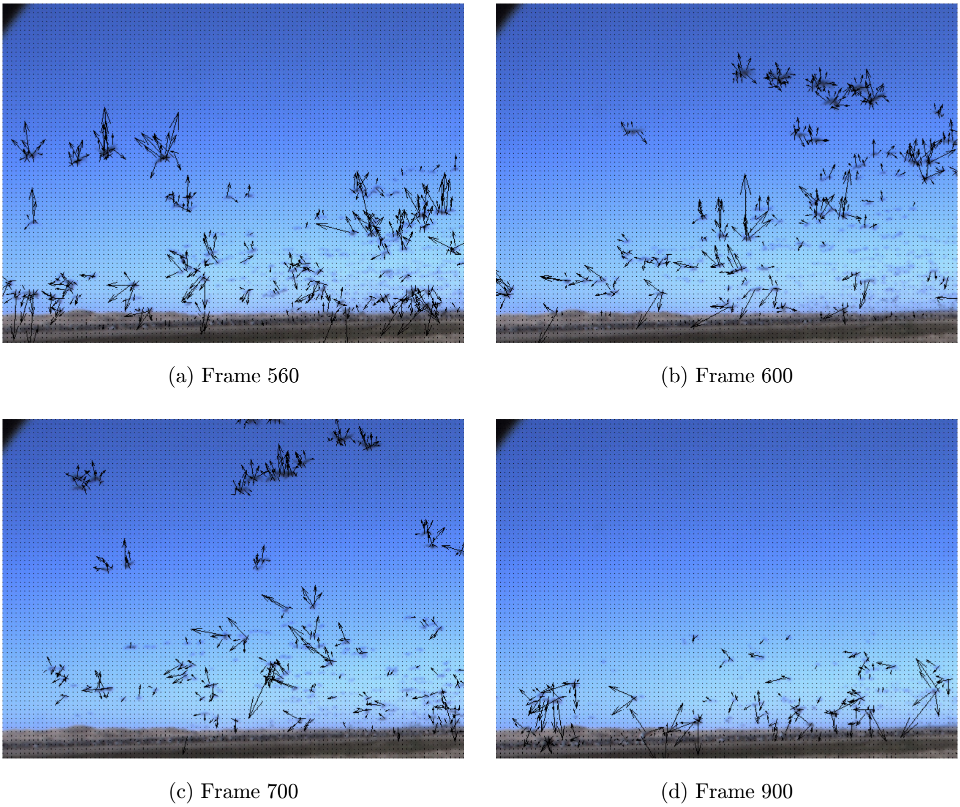
Fig. 5. Evaluation of video-based landing behavior recognition. It is seen that from approximately frame 600 to frame 850, flushing behavior is most likely. This is because some geese are flying towards the camera, resulting in upwards movement, and some geese are taking off to find other places to land.

and upward directed movement and a measure of the variance in the directionality of the optical flow estimates could be used to improve this in the algorithm. Figure 5 shows the probability output from the video-based classifier, when landing behavior is present. It is seen that at approximately frame 600, the probability of flushing behavior becomes greater than the probability of landing behavior. Four frames from this observation is shown in Fig. 6, where Figs. 6(b) and 6(c) are falsely classified.

The fusion slightly degraded the performance of recognition of flushing behavior, w.r.t. sensitivity, compared to video-based classification. This is because the soft output from the audio-based classifier tends to behave as a crisp output. This is seen in Fig. 7 where the probability of flushing behavior for both the audio- and video-based classifier, is plotted. When the probability of flushing behavior is high, it is close to one, and when it is low, it is close to zero. One way of dealing with this issue, could be to incorporate a bound on the weights for the individual classifiers. This should be done based on the performance of the individual classifier. This has not been done here, but could be investigated, when more audio data is available.

The performance of audio-based recognition of flushing behavior show poor performance with low mean value and a high standard deviation. As described in Sec. 2.3, the duration of audio flushing data is very short compared to the other behavior classes, and few misclassifications has a high impact on the performance measure. More flushing data would provide a more realistic evaluation of the audio-based flushing behavior, however it is worth mentioning that the accuracy is high (Table 3).

The data we used in this paper were chosen to ensure that the geese were visible to the camera, which influenced the amount of flushing data available for the classifiers.


(a) Frame 560

(b) Frame 600

(c) Frame 700

(d) Frame 900

Fig. 6. Frames from landing behavior. A subsampled optical flow estimate is plotted to show the direction and magnitude of movement in the frames. In frames 600 and 700, the video-based classifier assigns this behavior to flushing behavior, as most major velocity vectors are directed upwards.

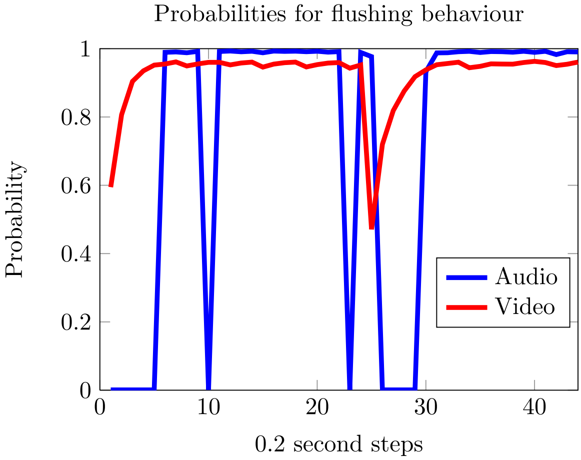


Fig. 7. Probability of flushing behavior.

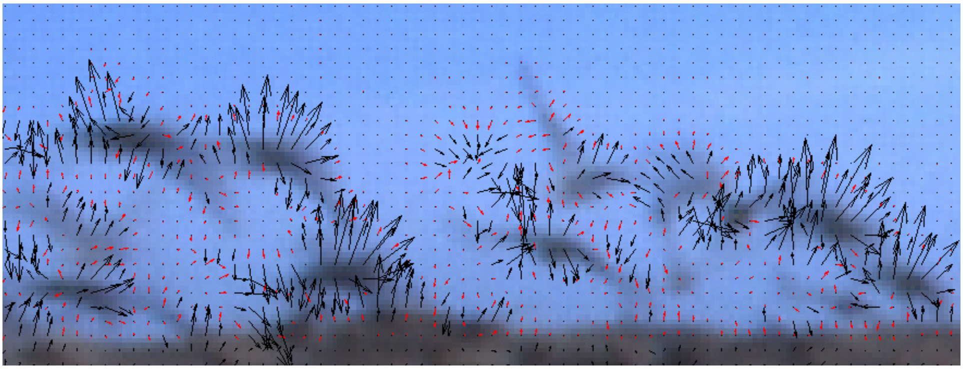


Fig. 8. The flow of behavior recognition.

Cameras have a limited field of view, and further research within this framework could include the use of omni-directional cameras, or temporal information to detect flushing behavior even after the geese have left the FOV of the camera, but still being within the range of the microphone.

When the geese are landing or flushing, the constant flapping of wings, produce optical flow vectors, which are directed both upwards and downwards. However, due to an image smoothing filter in the video preprocessing step and a threshold of the velocity vectors, this effect is reduced. In Fig. 8, this effect is shown in the case of flushing behavior. Here, the geese are very close to the camera, and the flapping of the wings introduce somewhat large velocity vectors. In the figure, the black arrows are the thresholded optical flow estimates (subsampled for illustration purposes). The red arrows indicate the smaller velocity vectors, which are removed due to thresholding. It is seen that some of the larger velocity vectors have a downwards directionality, due to e.g. the flapping wings. However, the video-based classifier is able to recognize the flushing behavior, as most of the larger velocity vectors originates from the goose and not its wings.

The video-based recognition depends on dividing the image into sky level and ground level. In this paper, this is done manually. However, in a real-life scenario, the camera position could be altered, deliberately or not. This would degrade performance, if the algorithm could not detect these levels automatically. Further work on the video-based recognition should therefore investigate if methods such as horizontal edge detection, hough transform or image clustering[17] could be utilized to accomplish this.

Furthermore, more adaptive or complex strategies for reliability ratio estimation exist, and more information about the signals, including signal-to-noise ratio and time of day, could be used to scale the outputs of the classifiers. The methods and framework presented in this paper makes the addition of new information straightforward. This includes using other classifiers for the single stream classification, given they output soft outputs or can be modified to do this.

8. Conclusion

Audio-visual fusion has been used for recognition of goose flocking behavior. The fusion of the audio- and video-based classifier has improved the recognition of goose flocking behavior. The improvement of using classifier fusion is most evident in the case of flushing and landing behavior recognition, where it was possible to combine the advantages of both the audio- and video-based classifier.

The improvement of landing behavior recognition is an important result in this research, since robust recognition of landing behavior is a critical component of an adaptive wildlife management system. Immediate detection of landing behavior is crucial to scare off geese while they are alert. Audio-visual recognition of goose flocking behavior may therefore potentially contribute to the reduction of goose related crop damage levels.

References

1. J. L. Barron, D. J. Fleet and S. S. Beauchemin, Performance of optical flow techniques, *Int. J. Comput. Vis.* **77** (1994) 43–77.
2. C. M. Bishop, *Pattern Recognition and Machine Learning* (*Information Science and Statistics*) (Springer-Verlag New York, Inc., 2006).
3. J. C. Brown and P. Smaragdis, Hidden Markov and Gaussian mixture models for automatic call classification, *J. Acoust. Soc. Am.* **125**(6) (2009) EL221–EL224.
4. W. M. Campbell, Speaker verification using support vector machines and high-level features, *IEEE Trans. Audio, Speech Lang. Process.* **10**(7) (2007) 2085–2094.
5. C. C. Chang and C. J. Lin, LIBSVM: A library for support vector machines (2011) http://www.csie.ntu.edu.tw/cjlin/papers/libsvm.pdf, pp. 1–39.
6. P. J. Clemins, M. B. Trawicki, K. Adi and M. T. Johnson, Generalized perceptual features for vocalization analysis across multiple species, *ICASSP* (2006), pp. 253–256.
7. B. N. Courty and T. Corpetti, Crowd motion capture, *Comput. Animations Virtual Worlds* **18** (2007) 361–370.
8. S. Davis and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuosly spoken sentences, *IEEE Trans. Acoust., Speech, Signal Process.* **28**(4) (1980) 357–366.
9. M. S. Dawkins, H.-J. Lee, C. D. Waitt, and S. J. Roberts, Optical flow patterns in broiler chicken flocks as automated measures of behaviour and gait, *Appl. Animal Behav. Sci.* **119**(3–4) (2009) 203–209.
10. J. R. Deller Jr., J. G. Proakis and J. H. Hansen, *Discrete Time Processing of Speech Signals* (Prentice Hall PTR, 1993).
11. L. P. Dmitrieva and G. Gottlieb, Influence of auditory experience on the development of brain stem auditory-evoked potentials in mallard duck embryos and hatchlings, *Behav. Neural Biol.* **61** (1994) 19–28.
12. E. Erzin, Y. Yemez and A. M. Tekalp, Multimodal speaker identication using an adaptive classifier cascade based in modality reliability, *IEEE Trans. Multimedia* **7**(5) (2005) 840–852.
13. S. Fagerlund, Bird species recognition using support vector machines, *EURASIP J. Adv. Signal Process.* **2007** (2007) 1–9.
14. D. J. Fleet and Y. Weiss, Optical flow estimation, in *Mathematical Models in Computer Vision*, eds. Paragios, N. Chen, Y. and Faugeras, O. (Springer, 2005), pp. 239–258.
15. T. Ganchev, N. Fakotakis and G. Kokkinakis, Comparative evaluation of various MFCC implementations on the speaker verification task, in *Proc. SPECOM*, (2005), pp. 191–194.

16. J. M. Gilsdorf, S. E. Hygnstrom and K. C. Vercauteren, Use of frightening devices in wildlife damage management, *Integr. Pest Manag. Rev.* **1**(27) (2002) 29–45.
17. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd edn. (Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 2006).
18. B. K. P. Horn and B. G. Schunck, Determining optical flow, *Artif. Intell.* **17**(1) (1981) 185–203.
19. E. Hu, T. Tan, L. Wang and S. Maybank, A survey on visual surveillance of object motion and behaviors, *IEEE Trans. Syst., Man Cybern. C, Appl. Rev.* **34**(3) (2004) 334–352.
20. J. Kittler, M. Hatef, R. P. W. Duin and J. Matas, On combining classifiers, *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(3) (1998) 226–239.
21. K. Kryszczuk, J. Richiardi, P. Prodanov and A. Drygajlo, Reliability-based decision fusion in multimodal biometric verification systems, *EURASIP J. Adv. Signal Process.* **2007** (2007) 9.
22. C. Lee, C. Chou, C. Han and R. Huang, Automatic recognition of animal vocalizations using averaged MFCC and linear discriminant analysis, *Pattern Recogn. Lett.* **27**(2) (2006) 93–101.
23. D. Magee, Detecting lameness using Re-sampling Condensation and multi-stream cyclic hidden Markov models, *Image Vision Comput.* **20**(8) (2002) 581–594.
24. G. Manteuffel, Vocalization of farm animals as a measure of welfare, *Appl. Animal Behav. Sci.* **88**(1–2) (2004) 163–182.
25. G. Manteuffel and P. C. Schön, Measuring pig welfare by automatic monitoring of stress calls, *Bornimer Agrartechnische Berichte* **29** (2002) 110–118.
26. P. Martiskainen, M. Järvinen, J.-P. Skön, J. Tiirikainen, M. Kolehmainen and J. Mononen, Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines, *Appl. Animal Behav. Sci.* **119**(1–2) (2009) 32–38.
27. G. Medioni, I. Cohen, F. Brémond, S. Hongeng and R. Nevatia, Event detection and analysis from video streams, *IEEE Trans. Pattern Anal. Mach. Intell.* **23**(8) (2001) 873–889.
28. T. A. Messmer, The emergence of human-wildlife conflict management: Turning challenges into opportunities, *Int. Biodeteri. Biodegr.* **45** (2000) 97–102.
29. D. Moura, W. Silva, I. Naas, Y. Tolon, K. Lima and M. Vale, Real time computer stress monitoring of piglets using vocalization analysis, *Comput. Electron. Agr.* **64**(1) (2008) 11–18.
30. D. Nolte, Behavioral approaches for limiting depredation by wild ungulates, in *Grazing Behavior of Livestock and Wildlife*, eds. Launchbaugh, K. Sanders, K. and Mosley, J. (University of Idaho, 1999), pp. 60–69.
31. J. C. Platt, N. Cristianini and J. Shawe-Taylor, Large margin DAGs for multiclass classification, *Adv. Neural Inform. Process. Syst.* **12**(3) (2000) 547–553.
32. G. Potamianos, C. Neti, L. Juergen and I. Matthews, Audio-visual automatic speech recognition: An overview, *Issues in Visual and Audio-Visual Speech Processing*, eds. Bailly, G. Vatikiotis-Bateson, E. and Perrier, P. (MIT Press, 2004).
33. D. Reby, R. Andre-Obrecht, A. Galinier, J. Farinas and B. Cargnelutti, Cepstral coefficients and hidden Markov models reveal idiosyncratic voice characteristics in red deer (Cervus elaphus) stags, *J. Acoust. Soc. Am.* **120**(6) (2006) 4080–4089.
34. K. A. Steen, H. Karstoft and O. Green, A multimedia capture system for wildlife studies, *The Third Int. Conf. Emerging Network Intelligence*, Lisbon, Portugal, November 2011, pp. 131–134.
35. K. A. Steen, O. R. Therkildsen, H. Karstoft and O. Green, A vocal-based analytical method for goose behaviour recognition, *Sensors* **12**(3) (2012) 3773–3788.

36. K. A. Steen, O. R. Therkildsen, H. Karstoft and O. Green, Video-based detection of goose behaviours, *Agriculture and Engineering for a Healthier Life*, Valencia, Spain, July 2012, p. 6
37. R. E. Thomas, K. M. Fristrup and P. L. Tyack, Linking the sounds of dolphins to their locations and behavior using video and multichannel acoustic recordings, *J. Acoust. Soc. Am.* **112**(4) (2002) 1692.
38. R. Tillett, Using model-based image processing to track animal movements, *Comput. Electron. Agr.* **17**(2) (1997) 249–261.
39. V. M. Trifa, A. N. G. Kirschel, C. E. Taylor and E. E. Vallejo, Automated species recognition of antbirds in a Mexican rainforest using hidden Markov models, *J. Acoust. Soc. Am.* **123**(4) (2008) 2424–2431.
40. E. E. Vallejo and C. E. Taylor, Adaptive sensor arrays for acoustic monitoring of bird behavior and diversity: Preliminary results on source identification using support vector machines, *Artif. Life Robot.* **14**(4) (2009) 485–488.
41. L. Wang, Recent developments in human motion analysis, *Pattern Recogn.* **36**(3) (2003) 585–601.
42. Z. Zeng, M. Pantic, G. Roisman and T. S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(1) (2009) 39–58.

Kim Arild Steen, born in 1984, holds a B.Sc. degree in Electrical Engineering ('09) and an M.Sc. degree in Technical Information Technology ('11) from Aarhus School of Engineering, Denmark. In 2011, he started his Ph.D. research on signal processing and pattern recognition methods for solving conflicts between agriculture and wildlife. During this research, he studied at the Georgia Institute of Technology, Center for Image and Signal Processing. His research interests include signal processing, image processing and pattern recognition.

Ole Roland Therkildsen, born in 1971, received his M.Sc. degree from the Department of Ecology and Genetics, Aarhus University. He is a research biologist at the Department of Bioscience, Aarhus University. As an expert in wildlife ecology and management, his research interests include human–wildlife conflicts, wildlife damage, waterbird feeding ecology and radar studies of migratory birds.

Ole Green, born in 1978, holds a B.Sc. degree in Mechanical Engineering from Vitus Bering, Horsens Denmark ('04), an M.Sc. degree in Agronomy ('06) from The Royal Veterinary and Agricultural University, Denmark and a Ph.D. in Biosystems Engineering ('10) from the University of Aarhus, Denmark. Since 2009 he has held various positions in academia and industry, including Head of research group — Automation and System Technology at the Department of Engineering, Aarhus University. Currently he is Head of Strategic Development at Kongskilde Industries (agricultural machinery development and production). His main research interest is the implementation of sensor technology and information systems on agricultural machinery for improved decision support systems.

Henrik Karstoft, born in 1962, holds a B.Sc. degree in Physics and in Computer Science, an M.Sc. degree in Math ('88) and a Ph.D. in Math (Differential Topology) ('91) from the University of Aarhus, Denmark. Since 2007 he has been Docent at the Aarhus School of Engineering, Aarhus University, where he leads the group in Signal Processing. Karstoft has experience in working on R&D projects in Image and Signal Processing in a professional R&D organization. His main interest is in applications of Image and Signal Processing. Currently his main interest is in applications of video tracking, pattern recognition and machine learning, in agricultural applications. He is also working on Image and Video enhancement and Reconstruction in Industrial Applications. Karstoft has published papers on Signal and Image Processing and Mathematics. Karstoft has been supervising four Ph.D. students, supervised more than 40 students in their M.Sc. thesis projects and more than 65 bachelor student's final projects. He is a member of the external examiner corps in Engineering and external examiner corps in Mathematics in Denmark and has been a member of several Educational Committees for E.E. and ICT studies at Aarhus School of Engineering, Aarhus University.

Copyright of International Journal of Pattern Recognition & Artificial Intelligence is the property of World Scientific Publishing Company and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.