Counting Large Flocks of Birds Using Videos Acquired with Hand-Held Devices

Amanda Dash and Alexandra Branzan Albu(✉)

Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada
{adash42,aalbu}@uvic.ca

Abstract. Due to the rapidly increasing quality of cameras and processing power in smartphones, citizen scientists can play a more significant role in environmental monitoring and ecological observations. Determining the size of large bird flocks, like those observed during migration seasons, is important for monitoring the abundance of bird populations as wildlife habitats continue to shrink. This paper describes a pilot study aimed at automatically counting birds in large moving flocks, filmed using hand-held devices. Our proposed approach integrates motion analysis and segmentation methods to cluster and count birds from video data. Our main contribution is the design of a bird counting algorithm that requires no human input, and functions well for videos acquired in non-ideal conditions. Experimental evaluation is performed using ground truth of manual annotations and bird counts, and shows promising results.

Keywords: Motion analysis · Tracking · Environmental monitoring · Bird counting

1 Introduction

Computer vision technologies can play an important role in environmental monitoring. Significant theoretical advances have been made recently in terms of automatic, fast, and reliable object detection, classification, and tracking. These advances enable the design of accurate methods for environmental monitoring applications [2] using a large variety of video data. Computer vision algorithms, combined with new image acquisition technologies, such as those using unmanned aerial videos [6], can trigger a revolution in the way wildlife monitoring surveys are performed.

Data acquisition for environmental surveys of large bird populations, such as bird flocks, can be challenging. A solution proposed for an airport-based bird surveillance system uses fixed thermal cameras in order to track flying birds [9]. A static system might work well for preventing bird strikes at an airport, but not for estimating bird abundance in large moving flocks. Bird flocks can be filmed by amateurs using standard cameras in mobile devices; however, the quality of

J. Blanc-Talon et al. (Eds.): ACIVS 2017, LNCS 10617, pp. 468–478, 2017.
https://doi.org/10.1007/978-3-319-70353-4_40

the video can vary considerably due to environmental conditions, background complexity, and camera quality.

Ornithologists have been surveying flocks and their moving patterns long before cameras and binoculars were common-place. Manual surveys are, however, time-consuming, and prone to human error; they are usually performed well only by few highly trained specialists. As there has been an increase in the number of bird reserves, partly due to the larger human impact on animal habitats in recent years, the need for more frequent and accurate bird population surveys has increased. Periodic bird surveys allow for tracking migrations patterns of various bird species. This information is vital for quantifying the effect of industrial development on the nature conservation value of the land and can also guide conservation policies to prevent the local, regional or global extinctions of birds with smaller population sizes or disappearing migration patterns. Manual bird counting from video data, even when assisted by computer programs, is a long, tedious, and error-prone process. This paper proposes an automatic method for estimating the number of birds in large flocks using videos acquired with hand-held mobile devices.

The remainder of the paper is organized as follows. Section 2 presents related work in environmental visual monitoring performed with computer vision methods and by human experts. Section 3 describes the proposed approach, and Sect. 4 discusses its experimental evaluation. Conclusions and future work directions are provided in Sect. 5.

2 Related Work

Computer vision methods are increasingly used to estimate diversity and abundance of animal populations. This is due in part to the ability of placing cameras in remote/inaccessible locations, which cannot be easily reached by human observers (i.e. deep sea and high elevation mountainous areas). Also, camera-based environmental observations have a low environmental impact. Methods for automatic fish classification [13,16] and counting [5] have been recently proposed. Fish and bird counting share similar challenges with regards to the automatic segmentation of the moving targets, mostly due to partial occlusions, poor visual conditions and relatively low video quality. Classification and counting of animal species and individuals from video data is a non-trivial task. Bird flocks are of higher density than schools of fish, and the size of individual birds is much smaller than the size of fish; thus, a straightforward adaptation of fish counting methods for bird counting purposes is impossible.

Tracking and/or counting based on prior knowledge of individual behaviour is used in dense crowd modeling [15]. Rittscher et. al. [12] proposed a human-counting approach that models behaviour based on colour signature, template matching, and probabilistic estimation of foreground data. Individual tracks are clustered in group tracks for counting. This algorithm under-performs in high-density crowds. Ali and Shah [1] approach the same problem by considering a person as a set of particles that are affected by external factors. This type of particle behaviour can be also applied to bird flocks [8].

The standard method used by ornithologists for obtaining an unbiased measure of bird abundance employs visual observation with the naked eye, or with binoculars and a spatial sampling strategy [7]. This involves selecting and following the motion of a relatively small, quasi-rectangular sample region within the flock. The bird count is performed only within this region, which is assumed to be representative of the overall bird density of the flock. The bird count is then extrapolated by estimating how many sample region areas are contained within the flock. This method outputs an approximate count which is considered accurate enough for inferring population statistics. The method relies heavily on expert-made, ad-hoc decisions such as the location and size of the sample region, and the count extrapolation from the sample region to the entire flock. The method fails when dealing with flocks that have a high variance in bird density, which violates the basic assumption underlying sampling. Also, it is difficult to manually count high density flocks (i.e. when birds are close and partially overlapping). Manual bird count is also performed with difficulty in low visibility, hazy conditions, and for fast moving flocks.

3 Proposed Approach

While our proposed approach follows the same general strategy as the manual, expert-based bird count, it attempts to improve its accuracy in the following three ways. First, we automate the spatial sampling process, i.e. the detection of sample regions, to be further called subregions; we perform a complete partition of the flock in subregions. Second, instead of extrapolating the count from the sample region to the entire flock, we compute individual bird counts in each subregion; we thus allow for different motion patterns to exist within different subregions. Third, temporal information is considered by using video segments (or clips) of a moving flock to account for overlapping birds and flock shape changes. The bird count is averaged over a sequence of frames in order to obtain a better approximation of the true number of birds in the flock.

The proposed approach consists in two main processing modules (see Fig. 1), namely subregion partitioning (Sect. 3.1), and bird detection (Sect. 3.2).

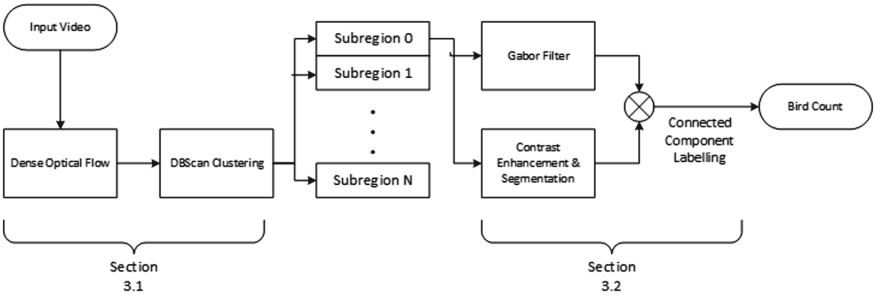


Fig. 1. Flow chart of the proposed method

3.1 Subregion Partitioning

For all frames composing the input video sequence, the dense optical flow algorithm [4] is used to find the region of interest (where birds are most likely to exist) by determining the motion flow vectors at uniformly sampled points in the frame. The motion flow vectors are partitioned into subregions by using the density-based spatial clustering clustering algorithm (DBScan) [3]. This method was chosen due to its high robustness with respect to noise. Each subregion is represented by the minimum bounding rectangle (MBR) of the clustered motion vectors.

Determination of the Region of Interest. In video sequences acquired using hand-held cameras, individual birds in a flock may be blurry due to camera motion, camera distance to flock, low light conditions, etc. Thus, the process of extracting features required for some optical flow algorithms, such as Lucas-Kanade [11], is not reliable. Instead, the Farneback [4] method is used; this method estimates the displacement field between two frames and attempts to compensate for background motion. We determine the region of interest by calculating the motion displacement field, $O(n)$:

$$F(x, y, n-1) = F(x + O_{\Delta x}(x, y, n), y + O_{\Delta y}(x, y, n-1), n) \tag{1}$$

where n is the frame index, (x, y) are spatial coordinates of the pixel, and $\Delta_{x,y}$ is the optical flow motion displacement between frames $F(n-1)$ and $F(n)$.

To reduce computation, only a subset of the motion vectors, $M(n)$, is used, as shown in Eq. 2a. This subset of motion vectors is used as the feature set to partition the region of interest into subregions using the DBScan clustering algorithm.

$$P(x, y) = p(x + O_{\Delta x}(x, y), y + O_{\Delta y}(x, y)) \tag{2a}$$

$$M(n) = \{P(x + \Delta_x, y + \Delta_y) \mid \Delta_{xy} > 0, x = 0, S, \ldots X, y = 0, S, \ldots, Y\} \tag{2b}$$

where p is the sampling point, S is the sampling rate, X is the frame width, and Y is the frame height. For our method, a spatial sampling rate of 10 pixels is used.

Motion Vector Clustering. The motion vector subset $M(n)$ is used to partition the region of interest into subregions. To obtain the subregions partition, $M(n)$ is clustered using DBScan [3]. The DBSCAN algorithm requires 2 parameters: *epsilon*, the Euclidian distance threshold, which specifies how close points should be to each other to be considered a part of a cluster; and $minPts$, the minimum cluster size. We assume that motion vectors that differ in magnitude by more than the sampling rate are not generated by the same "group" of birds moving in a consistent direction and speed within the flock. Therefore, we set the Euclidean distance threshold, *epsilon*, to the sampling rate (10 pixels). The minimum cluster size $MinPts$ is set to 3 points, due to the fact that this is the

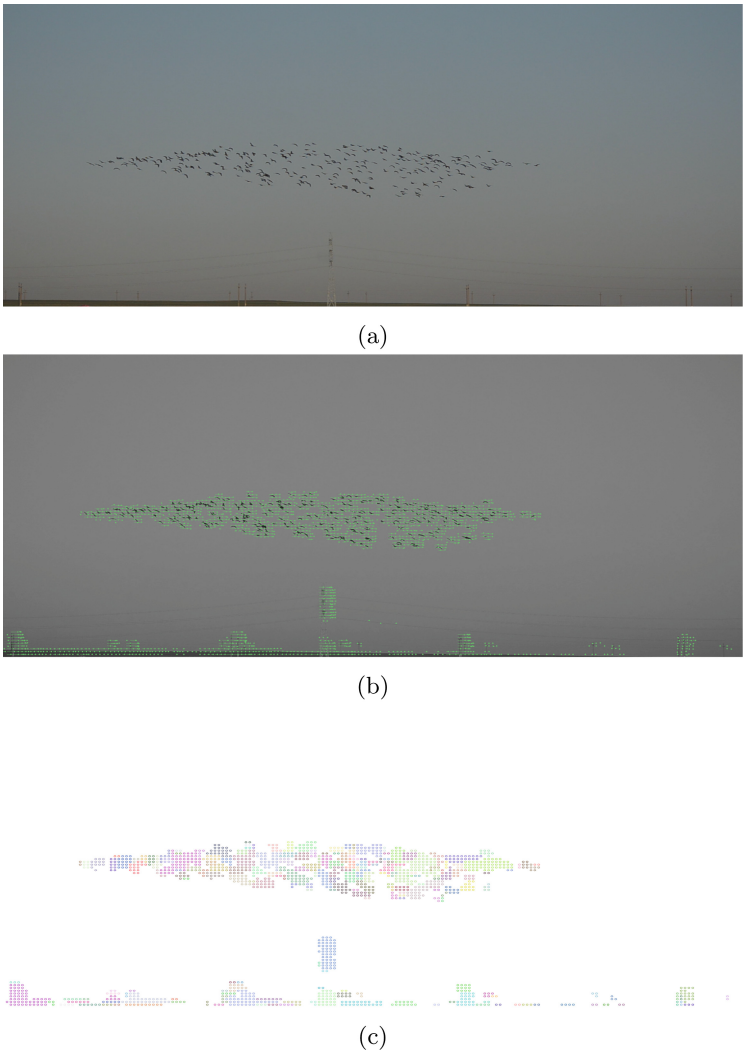


Fig. 2. Example showing how subregion partitioning works on frame $n = 64$ of video sequence $VID4$ in our dataset (a) Input frame (b) Motion displacement field $O(n)$, and (c) Result of motion vector clustering.

minimum number of points needed to define a planar surface. For each cluster i, its corresponding subregion $F_i(n)$ is found by taking the minimum bounding rectangle (MBR) of the cluster.

Figure 2 illustrates how subregion partitioning works on a typical frame from our experimental dataset.

3.2 Bird Detection

Videos of moving flocks of birds may exhibit variable degrees of blurriness, due to low contrast between birds and background (sky, grass, ground etc.) To reduce the number of missed bird detections, we perform edge enhancement using a non-photo-realistic rendering (NPR) algorithm [10], which exaggerates low-gradient edges via a palette reduction method.

The NPR algorithm performs a three-iteration loop on $F_i(n)$, using Eq. 3 to obtain a contrast-enhanced gray-scale image, $CE_i(n)$.

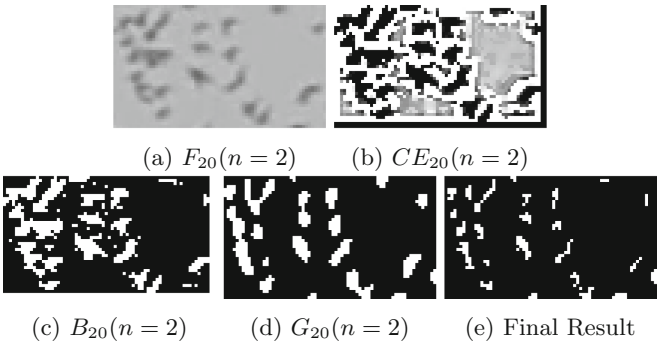
$$CE_i(x, y, n) = F_i(x, y, n) + \Delta F \tag{3}$$

where ΔF is computed using the intensity gradient between $F_i(x, y, n)$ and its' 5×5 neighbourhood.

Next, a simple thresholding operation is used to obtain the binary images $B_i(n)$ of the subregions, as follows.

$$B_i(n) = \begin{cases} 255 & \text{if } O_i(x, y, n) <= T_B \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where T_B is the threshold value, which was set to 10 based on experiments.



(a) $F_{20}(n = 2)$ (b) $CE_{20}(n = 2)$

(c) $B_{20}(n = 2)$ (d) $G_{20}(n = 2)$ (e) Final Result

Fig. 3. Example showing segmentation of subregion $i = 20$ in the second frame (n = 2) of video sequence $VID8$ in our dataset. The subregion i is contrast enhanced (b) and then thresholded (c). The binary result of applying Gabor filter is (d) The final result (e) is a pixelwise product of (c) and (d)

The adopted contrast enhancement method may introduce noisy structures that could be later on be falsely counted as birds. To compensate for this effect, a linear Gabor filter is applied to the subregion $F_i(n)$. The Gabor filter uses convolution in order to find the strongest vertical gradient regions, and aims to reduce the number of false bird detections.

The Gabor filter, G_i, is defined by:

$$\delta = 2.0/(s-1) \tag{5a}$$

$$i = k\delta cos(\theta) + m\delta sin(\theta) \tag{5b}$$

$$j = -k\delta sin(\theta) + j\delta cos(\theta) \tag{5c}$$

$$G_i(x,y) = e^{-\frac{(i^2+j^2)}{2\sigma^2}} \cos(2\pi i + 0.5\pi) \tag{5d}$$

where $\theta = 0$, s is the odd-kernel size, $x = \frac{s-1}{2} + i, y = \frac{s-1}{2} + j$, and $k, m \in \{-\frac{s-1}{2}, \frac{s-1}{2}\}$. The kernel size was set to 21.

The resultant convolution generates small connected regions where strong edges exist. This removes the low frequency textures like sky, grass and ground, but preserves the high frequency textures of the bird shape. The output of the Gabor filter is thresholded, resulting in the binary image $G_i(n)$.

To summarize, two binary images are produced via independent processes for each subregion, as follows: $B_i(n)$, resulting from contrast enhancement, and $G_i(n)$, resulting from Gabor filtering. The final result $SR_i(n)$ is the pixelwise product of $B_i(n)$ and $G_i(n)$. See also Fig. 3.

This image is then labelled using a standard single-pass 8-connected component labelling algorithm. To account for partially overlapping birds, the average area of all birds in the subregion, $\bar{a}$, is used to determine if there are connected birds that weren't separated during the bird segmentation.

The bird count per frame, $birds(n)$, is obtained using Eq. 6a.

$$birds_i = \sum \{ \lfloor \frac{A}{\bar{a}} \rfloor \mid a \in A \} \tag{6a}$$

$$birds(n) = \sum_{i \in SR} birds_i \tag{6b}$$

where SR is the number of sub-regions in frame n, and A is the set of all connected components in subregion $SR_i(n)$, and $\lfloor ... \rfloor$ is the floor operator. The summation term in Eq. 6a shows that two or more birds can be counted inside connected components with larger areas, i.e. areas that are greater than the average bird area.

Temporal information is considered in order to remove noise in the bird counts, which are performed on a frame by frame basis. For a given video, the average bird count over a sequence of frames of predefined length is computed as the final bird count (flock size) per video.

4 Experimental Evaluation

This section presents experimental results from a feasibility study for automatic bird counting from videos of bird flocks acquired with hand-held cameras. The purpose of this feasibility study is to identify which characteristics of input video data lead to accurate results from our proposed approach, as well as which characteristics of input video data result in failures for our approach. Our dataset

Table 1. Dataset video properties

VID	FPS	Resolution	Source	Conditions	Bird speed
1	30	1280×720	Youtube	Light clouds, clear	Medium
2	30	1280×720	Youtube	Horizon visible, sunset	Very fast
3	30	1920×1080	Youtube	Clear	Fast
4	24	1920×1080	SOR	Horizon visible, low light	Medium
5	30	1920×1080	Youtube	Light clouds, clear	Slow
6	24	1920×1080	SOR	Horizon visible, low light	Medium
7	30	1920×1080	Youtube	Cloud, sunset	Fast
8	24	1920×1080	SOR	Horizon visible, low light	Medium
9	30	1280×720	Youtube	Horizon visible, clear	Medium

consists of two types of video sequences. The first type is acquired by expert ornithologists affiliated with the Romanian Ornithological Society (SOR) [14]. The second type consists of video sequences of bird flocks that we have retrieved from YouTube. The assembly of the dataset attempted a holistic exploration of various environmental conditions, and speed of birds. All video sequences in our dataset were manually annotated for ground truth purposes.

4.1 Dataset Description

The experimental dataset consists of 9 video sequences with a duration of 3 s each, recorded using hand-held (non-stationary) devices. All video sequences verify three criteria: (a) flock is countable by human experts, (b) flock is scale invariant (no zooming), and (c) flock is not visually obstructed. For each video, frames were manually annotated with the location of each bird at half-second intervals for a total of 29,354 birds. The annotation process took approximately 15 h. Table 1 lists the video spatial and temporal resolution, environmental conditions, and qualitative assessment of bird speed.

4.2 Experimental Results

The proposed method was evaluated on an 64-bit i7 Linux desktop and had an average runtime of 1 min per second of video. The two main modules composing our method, namely bird segmentation and counting, were evaluated individually. This was done to provide more insight into sources of error introduced by two separate processes. Experimental evaluation uses standard metrics of detection tasks such as precision, recall, and average count error. Precision and recall are calculated for both the intermediate bird segmentation and the bird counting tasks. The birds segmentation outputs subregions (i.e. grids) that correspond to birds. To calculate the precision_{grid} and recall_{grid}, the following definitions were used:

$true\text{-}positive_{grid}$: the subregion i is correctly labelled as containing birds
$false\text{-}positive_{grid}$: the subregion i is incorrectly labelled as containing birds
$false\text{-}negative_{grid}$: the subregion i is incorrectly labelled as not containing birds

For the bird counting evaluation, the *true-positives* are the number of correctly labelled birds found in the evaluated frame, the *false-positive* is the number of incorrectly labelled birds, and the *false-negative* is the number of birds not detected by the proposed method.

Table 2. Algorithm results aggregated per video and compared to the average of the ground truth labelling. The precision and recall are calculated for the bird segmentation and bird counting tasks

VID	Ground truth	Proposed method	Count error	Precision_{grid}	Recall_{grid}	Precision_{bird}	Recall_{bird}
1	81 ± 0	93 ± 5	12	0.9175	0.9808	0.7915	0.9847
2	185 ± 29	269 ± 38	84	0.7471	0.8247	0.485	0.8508
3	286 ± 8	262 ± 48	-24	0.8916	0.7855	0.8289	0.8317
4	295 ± 6	322 ± 37	27	0.7769	0.9097	0.785	0.9097
5	535 ± 3	554 ± 26	19	0.9214	0.9191	0.9051	0.9358
6	607 ± 63	550 ± 82	-57	0.8147	0.8856	0.8315	0.8143
7	961 ± 37	1006 ± 102	45	0.8105	0.9103	0.799	0.9363
8	1154 ± 17	1050 ± 59	-104	0.8752	0.9031	0.7867	0.6354
9	786 ± 35	777 ± 58	-9	0.8706	0.8242	0.7463	0.8848
Average			**$\pm$42.3**	**0.8473**	**0.8826**	**0.7732**	**0.8648**

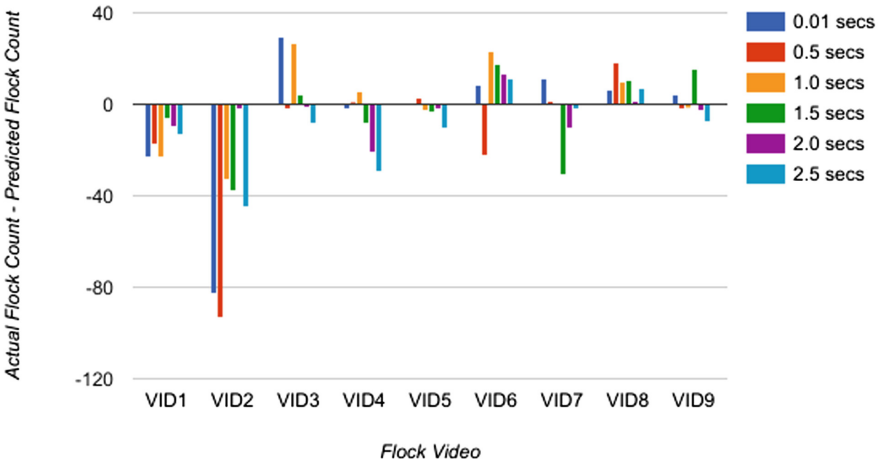


Fig. 4. The actual bird counts vs the predicted flow counts for all videos in the sample dataset.

We compared our approach against the ground-truth labelling at the uniformly sampled temporal moments. The average values over the 6 sampled

moments (at half-second intervals) for each video are aggregated in Table 2. The average miscount was 42.3 birds per video, with the worst average count error of -104 belonging to VID8. Each video had a tendency either towards under- or over-counting, as shown in Fig. 4. The best performer was VID5, which had a low average count error and high precision and recall. This is most likely due to the low speed of birds in VID5, which resulted in less blurring between the birds and background.

The worst performers were on VID7 and VID8 which had a miscount of -284 and $+210$, respectively. Since temporal evaluation moments were chosen via uniform sampling at every 0.5 s, no additional image analysis was performed to evaluate the viability of the frame selected for the count. VID7 performed poorly at 1.5 s because the birds were blurry due to either flock movement or camera movement. This caused the segmentation to group too many birds together, resulting in severe under-counting. In VID8, the flock of birds was close to the ground and the algorithm mislabelled parts of the field as birds Precision_{bird} for VID2 is very low, only 0.485; this video contained very fast moving birds, and suffered from the same segmentation problems as VID7 as a result. One may note that, for each analyzed video, a selection of optimal frames for bird counting and evaluation purposes, based on image quality criteria, is likely to improve results.

5 Conclusion

This paper proposes an automatic bird counting approach using videos acquired with hand-held devices. Based upon preliminary evaluation results, our proposed approach is likely to have a significant impact on future surveys of abundance and migratory patterns in various bird populations. Our proposed method performed similarly well for all videos with respect to various environmental conditions and flock sizes. It is very encouraging to conclude that, as the flock size grew, the relative error in bird counts remained roughly unrelated to the size of the flock; this is definitely a strong advantage with respect to manual count. There is a strong correlation between the performance of the proposed method and the quality of the frame used for evaluation. Incorporating more temporal information to avoid false detection and partial bird overlaps, as well as automatic frame selection based on quality criteria will likely yield improved performance.

References

1. Ali, S., Shah, M.: Floor fields for tracking in high density crowd scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008. LNCS, vol. 5303, pp. 1–14. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88688-4_1
2. Dell, A.I., Bender, J.A., Branson, K., Couzin, I.D., de Polavieja, G.G., Noldus, L.P., Pérez-Escudero, A., Perona, P., Straw, A.D., Wikelski, M., et al.: Automated image-based tracking and its application in ecology. Trends Ecol. Evol. **29**(7), 417–428 (2014)

3. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD. vol. 96, pp. 226–231 (1996)
4. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 363–370. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-45103-X_50
5. Fier, R., Albu, A.B., Hoeberechts, M.: Automatic fish counting system for noisy deep-sea videos. In: Oceans-St. John's 2014, pp. 1–6. IEEE (2014)
6. Gonzalez, L.F., Montes, G.A., Puig, E., Johnson, S., Mengersen, K., Gaston, K.J.: Unmanned aerial vehicles (UAVs) and artificial intelligence revolutionizing wildlife monitoring and conservation. Sensors **16**(1), 97 (2016)
7. Gregory, R.D., Gibbons, D.W., Donald, P.F.: Bird census and survey techniques. Bird Ecol. Conserv., pp. 17–56 (2004)
8. Hartman, C., Benes, B.: Autonomous boids. Comput. Anim. Virtual Worlds **17**(3–4), 199–206 (2006). https://doi.org/10.1002/cav.123
9. Huang, Y., Zheng, H., Ling, H., Blasch, E., Yang, H.: A comparative study of object trackers for infrared flying bird tracking. arXiv preprint arXiv:1601.04386 (2016)
10. Li, H., Mould, D.: Contrast-enhanced black and white images. In: Computer Graphics Forum, vol. 34, pp. 319–328. Wiley Online Library (2015)
11. Lucas, B.D., Kanade, T., et al.: An iterative image registration technique with an application to stereo vision. In: International Joint Conference on Artifical Intelligence (IJCAI), vol. 2, pp. 674–679 (1981)
12. Rittscher, J., Tu, P.H., Krahnstoever, N.: Simultaneous estimation of segmentation and shape. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005, CVPR 2005, vol. 2, pp. 486–493. IEEE (2005)
13. Rodrigues, M.T., Freitas, M.H., Pádua, F.L., Gomes, R.M., Carrano, E.G.: Evaluating cluster detection algorithms and feature extraction techniques in automatic classification of fish species. Pattern Anal. Appl. **18**(4), 783–797 (2015)
14. Romanian Ornithological Society, (2016). www.sor.ro
15. Saleh, S.A.M., Suandi, S.A., Ibrahim, H.: Recent survey on crowd density estimation and counting for visual surveillance. Eng. Appl. Artif. Intell. **41**, 103–114 (2015)
16. Spampinato, C., Giordano, D., Di Salvo, R., Chen-Burger, Y.H.J., Fisher, R.B., Nadarajan, G.: 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, ARTEMIS 2010, ACM, New York, NY, USA, pp. 45–50 (2010). http://doi.acm.org/10.1145/1877868.1877881