IEICE Technical Report
IMQ2021-21,IE2021-83,MVE2021-50(2022-03)

Towards Detecting Birds from Panorama Video Aided by Sound Source Localization

Baidong CHU†, Chihaya MATSUHIRA†, Yasutomo KAWANISHI††,†, Marc A. KASTNER†††,†, Takahiro KOMAMIZU††††, Ichiro IDE†,††††, and Daisuke DEGUCHI†

† Graduate School of Informatics, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464-8601 Japan
†† Guardian Robot Project, RIKEN
2-2-2 Hikaridai, Seika-cho, Souraku-Gun, Kyoto, 619-0288 Japan
††† Digital Content and Media Sciences Research Division, National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
†††† Mathematical and Data Science Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464-8601 Japan
E-mail: † {chubd, matsuhirac}@cs.is.i.nagoya-u.ac.jp, ddeguchi@nagoya-u.jp
†† yasutomo.kawanishi@riken.jp
†††mkastner@nii.ac.jp
†††† taka-coma@acm.org, ide@i.nagoya-u.ac.jp

Abstract In this report, we study a method to detect birds from a panorama video aided by Sound Source Localization (SSL). In the video, birds are relatively tiny to be detected from panorama frames. In the proposed method, birds are roughly localized in audio data by SSL algorithms, then corresponding regions are cropped from video frames and input to a Convolutional Neural Network (CNN) for detection. By narrowing down the searching area with SSL, relatively tiny birds in large video frames can be detected, and both detection precision and time performance are improved. Finally, we applied our method to a bird dataset and confirmed its effectiveness.

Key words bird detection, sound source localization, audio-visual information

1. Introduction

Detecting and monitoring birds is essential in multiple research fields, such as natural environment protection, aircraft safety, and so on. In order to detect and monitor birds, many devices are developed for recording audio or video data of birds. Following, many methods for using this data in bird detection and monitoring are developed.

Gayk and Mennill [1] estimate the position of the vocalizations of birds on wings based on eight microphones. Vemeycken et al. [5] use an ever-larger dense array with 64 microphones to simultaneously localize vocalizing songbirds in a radius of 75 m. Sumitani et al. [2] use audio data for bird localization and behavior analysis based on Sound Source Localization (SSL) algorithms. In their research, they also collect both audio and video data of birds by specially designed devices, and release a software called HARKBird [3] for bird sound analysis.

For bird detection from images, Researchers have performed some works to detect birds in realistic environments with CNNs. Yoshihashi et al. [8] use CNN-based methods for bird detection on a dataset including birds at a wind farm. They also proposed a method that focuses on detecting small birds in landscape images by combining deep features for objects from several CNN structures [4].

However, in real-world monitoring environments, birds can be relatively tiny, making them hard to be detected from large video frames. Further, audio recordings of birds can have much noise that badly influences the detection. Therefore, it is necessary to eliminate such errors by combining the two modalities.

Inspired by research of bird detection using either audio or video data, we propose a combined method that can detect birds from a panorama video aided by SSL results. The proposed method narrows down the searching area by applying sn SSL algorithm and cropping the corresponding regions of

Table 1 Examples of HARKBird results.

Time [s]	**ID**	**Azimuth** [rad]
0.5	0	−0.17
0.5	1	2.96
1.0	0	−0.08
1.0	1	2.79
...	...	...

(a) Example of a panorama video frame (b) Angle α

Figure 3 Video frame and angle di erence α.

3.1 Bird Localization Using Audio Data

As the first step of the proposed method, bird locations are calculated with the help of HARKBird from audio data. HARKBird is a powerful tool that is developed for bird sound analysis. SSL for a audio data can be performed in HARKBird by setting parameters of the recording device.

Table 1 shows examples of detection results of HARKBird. The results include the time and the azimuth of each sound source. Results at the same time are assigned with an ID in order to distinguish them.

3.2 Angle-Pixel Matching and Video Cropping

An SSL result in HARKBird includes time and azimuth. For an SSL result at time t with an azimuth θ, corresponding regions are cropped as follows. Note that all angles below are specified in degrees.

At first, frames with a number in the interval $[(t-\frac{p}{2})\,f, (t+\frac{p}{2})\,f)$ are extracted for further processing, where f [fps] is the frame rate of the video, and p is the sampling period of the SSL algorithm. This can cover all frames without overlap. In this report, we set $f = 29.97$ and $p = 0.5$.

Then, the azimuth θ is converted to pixel coordinate. As shown in Fig 1(a), arrangement and directions of the camera-mic pairs are already known. Therefore, the conversion can be performed with the help of the center position of the camera-mic pairs, as highlighted by colored boxes in Fig. 3(a). By knowing the width w of the video frame, the pixel coordinate x_θ of a corresponding azimuth θ can be calculated as follows:

$$x_\theta = \left\lfloor \frac{\theta - \alpha}{360} w \right\rfloor \mod w, \tag{1}$$

where α is the angle di erence between camera-mic pair 1 and the edge of the frame, shown as the red arrow in Fig 3(b). In this dataset, α is measured to be 2.72°.

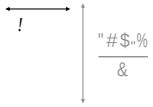


Figure 4 Example of a cropped region.

Next, a crop degree range φ is set, and frame regions between $[\theta-\varphi, \theta+\varphi]$ are cropped. θ is converted to x_θ according to Eq (1), and the degree φ can be converted to length l as follows:

$$l = w\,\frac{2\varphi}{360}. \tag{2}$$

As such, frame regions with width l centered on x_θ are cropped for further detection.

Finally, since HARKBird can only output lateral directions of birds as azimuths, the frames are cropped only in that direction. This results in a very slender region, which can have a bad influence on further detection. To solve this, cropped regions are further divided equally into v sub-regions in the vertical direction. The final size of a cropped region is $l\,\frac{h}{v}$ [pixels], where h is the height of the frame. An example of cropped regions with $v = 3$ is shown in Fig. 4.

3.3 Bird Detection from the Cropped Region

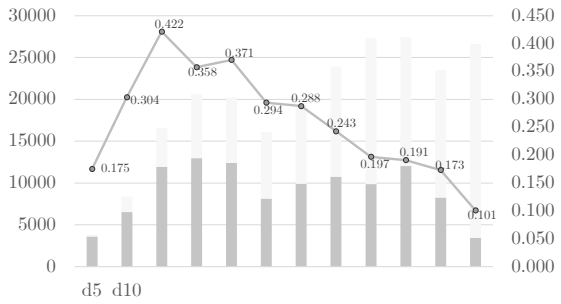
After cropping frame regions according to the SSL results, CNN models are used for the final bird detection. For this task, we compared several object detection models in Detectron2 model "zoo" [7]. Considering both time and precision performance, *Faster R-CNN_R101_FPN_3x* is chosen as the detection model.

4. Experiments

4.1 Experiment Settings

In the report, HARKBird 2.0.8(*3) is used for bird sound analysis. The Faster-RCNN model is deployed through Detectron2 and fine-tuned on the Caltech-UCSD Birds-200-2011 (CUB-200-2011) [6] dataset for 50 epochs with the Stochastic Gradient Descent (SGD) optimizer. The dataset is randomly divided to train, valid, and test sets with a 3:1:1 ratio. The learning rate starts with 10^{-3} and is multiplied by a scale factor of 0.1 at epoch 30 and epoch 40. The batch size is set to 64 in the fine-tuning process. Experiments are performed on a single NVIDIA RTX 3080 GPU and an AMD Ryzen 5950x CPU.

(*3) : https://sites.google.com/view/alcore-suzuki/home/harkbird/ (Accessed Feb. 12, 2022)



(a) Experiment on different crop range degrees

(b) Experiment on different vertical division numbers

Figure 5 Results of preliminary experiments.

Each method is evaluated by the following criteria:

- Average Precision (AP, main evaluation criterion)
- Number of True Positive (TP) and False Positive (FP) results with a confidence threshold of 0.1
- Detection time for all frames in the dataset

4.2 Preliminary Experiments on Parameter Selection

In the proposed method, the crop range degree φ and the vertical division number v need to be set for better precision and time performance. Due to this, different values are preliminary tested to find the best parameters of the proposed method.

In the first preliminary experiment, φ is set from 5° to 60° with an interval of 5°. Meanwhile, v is fixed to 3. Results of this experiment are shown in Fig. 5(a). We can see from the results that as φ increases, more FP results are detected because more noise is also input to the CNN model with a wider range. According to the AP, the best degree φ is decided as 15°.

In the second preliminary experiment, v is set from 1 to 5 to determine the optimal setting of v. φ is fixed to 15°, as the best value obtained in the first preliminary experiment. The results of this experiment are shown in Fig. 5(b). According to the results, the best number of vertical division v is decided as 3. We can also see that a division with an even number has a significant reduction in AP. A possible reason for this is that birds in the video data tend to appear near half the height of the frame. Therefore, with an even division number, birds tend to appear at upper or lower edges of the divided regions, which may have caused the low precision.

Table 2 Results of the main experiment.

Method	Time↓	AP↑	TP@0.1↑	FP@0.1↓
Raw frame (Baseline)	**0:39:31**	0.000	0	**3**
Div-12×3 (Comparison)	8:24:23	0.133	**14,705**	16,345
SSL-d15-div3 (Proposed)	1:53:52	**0.422**	11,947	4,637

4.3 Evaluation of the proposed method

In this experiment, the performance of several methods is compared and tested. In the first method, raw frames are input to the CNN model for bird detection without cropping (Raw frame). In the second method, raw frames are divided into 12×3 smaller blocks (Div-12×3), corresponding to the best φ ($2\varphi = 30 = \frac{360}{12}$) and v in the preliminary experiments. Each block is then input to the CNN model for detection. In the last method, the proposed method with $\varphi = 15^\circ$ and $v = 3$ is also evaluated (SSL-d15-div3). Examples of different inputs of the three methods are shown in Fig. 6.

Results of the experiment are shown in Table 2. Due to the relatively tiny bird size in raw frames, no bird was detected in the baseline method. The Div-12×3 method yielded more TPs than the proposed method, but also a higher number of FPs, resulting in a low AP. In the proposed method, on the other hand, using SSL outputs for cropping the video frames significantly improved its AP. Examples of detection results are shown in Fig. 7

As for computational complexity, detection with raw frames was the fastest, but it could not detect any bird. The proposed method ran four times faster than the Div-12×3 method, as it omitted the unnecessary regions with the aid of the SSL outputs.

From the results, we can conclude that the proposed method improved the detection results both in time and precision by narrowing down the searching area according to SSL results.

5. Conclusion

In this report, we studied a method to detect relatively tiny birds in large video frames aided by SSL. In the proposed method, we narrowed down the searching area with SSL results. By applying our method to a bird dataset, we confirmed that our method can improve both precision and