Visual Object Detector for Cow Sound Event Detection

YAGYA RAJ PANDEYA, BHUWAN BHATTARAI, AND JOONWHOAN LEE
Division of Computer Science and Engineering, Jeonbuk National University, Jeonju 54896, South Korea
Corresponding author: Joonwhoan Lee (chlee@jbnu.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education under Grant 2019R1A6A1A09031717.

ABSTRACT Sound event detection (SED) is a reasonable choice in a number of application domains including cattle sheds, dense forests, or any dark environments where visual objects are usually concealed or invisible. This study presents an autonomous monitoring system based on sound characteristics developed for welfare management in large cattle farms. Two types of artificial audio datasets are prepared: the cow sound event dataset and the UrbanSound8K dataset, which are then used with various sound object detectors for real world implementation. Using a data-driven approach, a conventional convolutional neural network structure with certain improvements is first applied, and from there proceed to a two-stage visual object detection method for audio by treating acoustic signals as an RGB images. The object detection method achieves a higher quantitative evaluation score and more precise qualitative results than previous related studies. We conclude that visual object detection methods are more effective than currently-available CNN architectures for rare sound object detection. Indeed, an artificial data preparation strategy can provide a better method for addressing the problem of data scarcity and the annotation difficulties involved in rare sound event detection.

INDEX TERMS Rare sound event detection, synthetic dataset, visual object detection model, convolutional network.

I. INTRODUCTION

Cattle behavior analysis has been an important issue since the earliest human civilizations. The housing, feeding, and management of cattle determines their wellbeing. Such wellbeing contributes to their behavior and impacts their health and productivity. Compared to other grazing animals, cattle in restricted environment have relatively high stereotypic behaviors and it is therefore necessary to observe closely their behavior and take the necessary steps to ensure their health and wellbeing. Generally, cattle express their condition through their behavior and by producing sounds. The study of cattle’s wellbeing is a generally broad area of research and our focus here is exclusively on sound event detection (SED) for cattle in a restricted environment. In terms of sound, the psychology of animals can be considered as a form of sonic onomatopoeia that can be used for communication, anticipation, mating, warnings and threats, pain, stress, hunger, or more. SED and sound analysis is one research trajectory that is concerned with the variety of sounds produced by animals and their corresponding temporal location.

A. LITERATURE REVIEW

SED involves temporal localization and categorization. The task of localization is to estimate the onset and offset times of audio events based on the segmentation of frame-level class predictions. Categorization refers to the task of audio tagging, which involves identifying the presence or absence of acoustic events in an audio recording without knowing their precise location in a temporal sequence. The goal of acoustic event detection is to detect and identify sound events as they occur in an audio clip.

Convolutional neural networks (CNNs) [1], [2] have become the architecture of choice for SED. Among these, the recurrent neural network (RNN)-based SED method [3], [4] adopts a large temporal context and performs relatively better than basic CNN structure. Hybrid structures containing both CNN and RNN layers, known as convolutional recurrent neural networks (CRNNs) [5], [6] are being developed for SED that integrate both spatial and temporal properties of the audio signal. A CRNN for joint
sound event localization and detection (SELD) of multiple overlapping sound events in three dimensional (3D) space has been proposed by Adavanne et al. [7]. Region-based fully convolutional network architectures [8], [9] has been presented for SED in the DCASE 2017 [10] Challenge dataset. An SED approach using graph Laplacian regularization [11] has been presented for co-occurring sound events in a single sound clip.

B. PROPOSE OF STUDY
The automation of welfare management is necessary in order to increase both the number of cow farms and the number of livestock they contain. SED using sound onomatopoeia provides a way of understanding the condition of the animal producing the sound. Over the past decade, a number of theoretical studies [16], [17] on cow behavior analysis and welfare management have been performed. Practical studies of automated surveillance of cow behavior and the detection of anomalies have focused on the estrous cycle in cows [18], [19], particularly estrus (in heat) detection for prediction of the timing of insemination [20]. To the best of our knowledge, however, no data-driven study has yet been made that focuses specifically on the detection and classification of cow sound events. The present study thus seeks to contribute to research on the real-time monitoring of the acoustic behavior of cows in a controlled environment with the purpose of improving welfare management while minimizing cost.

C. CONTRIBUTION
This article first presents two artificial audio datasets using the events of the cowshed and urban areas. We collect four categories of cow sound events and embed them in a variety of acoustic environments. Using a data-driven approach, a conventional convolutional neural network structure with certain improvements is applied in stage one, and from there proceeded to a two-stage Video Object Description Models (VODMs) for audio by treating acoustic signals as RGB images [12]. Four well-known VODMs namely the faster region-based convolutional neural network (FRCNN) [13], feature pyramid network (FPN) [14], cascade R-CNN (CRCNN) [15] and cascade FPN were trained and tested in this experiment.

The efficiency of the proposed CNN-based architecture is tested on three real and artificial unseen datasets. The proposed CNN outperforms similar previous CNN methods, but the results are inadequate in a real-world implementation. Next, an artificial audio mix is processed as a three-channel image and VODMs for SED are applied using Mel-spectrogram as input. For the artificial data, audio augmentation is applied to acoustic events to reduce overfitting and improve generalization. An event onset point is randomly selected by the avoidance of overlapping audio signals on an artificial audio mix. Analysis of a number of time-frequency representations is performed and section is based on those whose audio events are most perceptually distinguishable from the background scene. Cow sound events are computed along the temporal axis and system performance is evaluated using three test datasets.

The cow sound-related study is extended to a new rare sound event environment to investigate the robustness and diversity of the system. The aim of the study is to identify the best sound object detector for occurrences of rare sound events in a cowshed. We compare our CNN -research with a previous study [1] and our VODMs research with a previous study [12] of SED using the FRCNN detector. In comparison with previous studies, we obtain improved results by using a more advanced visual object detection network architecture and more effective data preparation techniques.

D. SYSTEM APPLICATIONS
Automatic SED and classification have a wide range of applications for animal behavior analysis and welfare management. SED is applicable to reducing the number of workers at a large cattle farm, to wildlife surveillance for behavioral analysis of endangered animals, and to animal or bird sound analysis in dark caves. Visual object detection fails in environments where objects of interest are either invisible or concealed, and only an acoustic signal is present. Another application for SED is as a semi-automatic annotation tool for fast and accurate annotation. A trained SED model can automatically filter sound events from a noisy environment and make quick evaluations of their location. System recommendations can assist in speeding up the annotation process even for noisy and extended audio recordings.

E. ORGANIZATION
The remainder of this article is organized as follows. Section II details the synthetic dataset preparation, including basic concepts and criteria, event and scene audio selection, and spectral representation. Section III explains the various system architectures and their use in sound event detection. Section IV presents the experimental results and corresponding analysis. Finally, Section V concludes the article.

II. DATASET PREPARATION
Real-world environments may include the presence of both monophonic and polyphonic sound events. While monophonic SED presupposes that events can be localized sequentially, the simultaneous occurrence of events, known as polyphonic SED, is common in everyday sound environments. This study, however, is concerned with monophonic SED in cowsheds and sound events in urban areas.

The limited quantity of annotation data is a major drawback when using a supervised data-driven approach. Although real-world recordings are easy to obtain and correspond closely with authentic scenarios, precise annotation is time-consuming and difficult to scale. Multiple instance learning and weakly-supervised learning provide semi-supervised methods for sound event detection. A number of studies [2], [21], [22] have been conducted for SED with weakly-labeled data in which the audio is tagged with
events without prior knowledge of the exact onset and offset temporal locations. Although this blind annotation strategy is an effective learning paradigm that addresses some of the difficulties of annotation, weakly-labeled datasets usually suffer from ambiguity problems due to background noise, cross noise, and fade in and fade-out effects. In addition, the accuracy of systems trained using weakly-labeled dataset is usually lower than those that have been trained using strongly-labeled datasets.

Manual annotation of huge datasets is a costly task in terms both of time and effort. While manual annotation of sound-scene audio material can be performed relatively quickly, for sound events the process is much slower because determining temporal boundaries is a vague and subjective process. Synthetic data provide one way of overcoming the problem of limited data, and overcoming annotation difficulties for SED using deep neural network architectures. Synthetic data are usually generated by superimposing sound events onto corresponding natural scenes at any temporal location. Artificial onset and offset points and corresponding class labels can be saved automatically during data preparation. The main challenge of this approach is the domain adoption of events and artificial background scenes as real scenarios. The network may learn trivial shortcuts from the synthetic data and may not perform as expected on concealed samples. Accordingly, we implemented five criteria, listed below, during the preparation of synthetic datasets.

1. Synthetic audio must have high diversity in event samples and background noises.
2. Events and background scenes must correspond closely with authentic acoustic environments.
3. Artificial mixes must include variations in instances and scenes when generating a train-and-test dataset.
4. Event onset points must be chosen randomly when writing to scene audio.
5. Individual sound events must be clearly distinguishable from one another in multi-event synthetic audio.

In this study, two synthetic datasets for rare sound event detection (R-SED) were prepared by following these five guidelines. The synthetic datasets and related code generated are available on GitHub. A detailed explanation of the two synthetic datasets is provided in the following subsections.

A. COW SOUND EVENT DATASET

In everyday life, the cows commonly do not make specific sounds to express states such as hunger, being in heat, or pain. Detecting such anomalies is valuable information for managing cowsheds with low manpower support. Accurately annotating data is indispensable to providing accurate information for supervised training. In this experiment, cattle sound events were first collected from the Internet and cowsheds using multiple audio-enabled CCTV cameras. Events were of four types: cow/calf lowing sounds, bull bellowing sounds, and mixed sounds produced by several animals at the same time. Audio was recorded in cowshed via a mono channel at a sampling rate of 8kHz. Data was collected over five months, mostly in mornings and evenings when cattle produce sounds more frequently.

Background scenes for the synthetic dataset were obtained from online sources (3,276 samples) and from DCASE 2017 Task 2 [10] noise audio (1121 samples). The sounds obtained online were extracted from videos of cowsheds, urban areas, and grass fields (grazing cows) in order to simulate a realistic environment. All audio recordings of scenes were edited into 10-second segments and sound events with a maximum length of 3 seconds were pasted onto them. Event length was set according to maximum event duration, while the scene length was selected so that a maximum of three events could be included without overlapping. Lastly, 47,400 samples were synthetically generated comprising four sound categories: Bull, Calf, Cow, and Mix. Onset points were selected randomly for each event, with no overlapping of events in multi-event audio. Statistical details of cow sound events are provided in Table 1. The number of events in this synthetic dataset was kept balanced by the use of multiple audio augmentation methods as in [23]. A total of 1152 samples were also manually annotated to simulate real-world scenarios. The network architectures were trained to detect sound events on only the temporal axis by keeping the frequency axis constant.

The three test datasets were constructed to evaluate the strength and adaptability of the trained object-detection framework in terms of real-world applications. The first set (Set-I) included 600 synthetic data samples, while the second (Set-II) included a combination of synthetic and manually-annotated real data with 866 samples. The third set (Set-III) contained 900 real data samples obtained from online sources and cowsheds. All the real-world data samples were manually annotated and each sample was identical in each test set. A statistical representation of the events in the synthetic test dataset is provided in Table 1.

| TABLE 1. Cow sound event distribution in synthetic train and test dataset. |
|---|---|---|---|---|---|---|
| | Train events samples | Test events samples |
| | Bull | Calf | Cow | Mix | Bull | Calf | Cow | Mix |
| **2520** | **2864** | **2916** | **2968** | **150** | **150** | **150** | **150** |
| * Original plus augmented audio samples. |

B. URBAN_SOUND8K DATASET FOR SED

The UrbanSound8K dataset [24] consists of 8,732 labeled sound excerpts, with a maximum length of 4 seconds, organized into 10 classes. Among these, five target classes—car horn, dog bark, gun shot, jackhammer, and siren—were selected to provide a fair comparison with the method in Eventness. Since the dataset used in Eventness and related source code are not publicly available, we prepared a new artificial dataset using the same UrbanSound8K events with background scenes as in the cow sound event dataset.
described above in Section II-A. Statistical details of the class events used in our experiment are provided in Table 2. In total, 24,000 training samples and 2,000 test samples from five classes were generated synthetically with diversity in scenes, events, and onset points on the temporal axis. Each synthetic audio sample included either one or two randomly-selected events with a maximum length of 4 seconds. The two events were written without overlaps onto the artificial audio mix.

### Table 2. Event distribution in artificial Audio using UrbanSound8K dataset.

| Class event | Train | Test | Total train* |
|-------------|-------|------|--------------|
| Car horn    | 344   | 85   | 1600         |
| Dog bark    | 800   | 200  | 1600         |
| Gun shot    | 300   | 74   | 1500         |
| Jackhammer  | 800   | 200  | 1600         |
| Siren       | 744   | 185  | 1488         |

* Original plus augmented train audio samples.

### C. SPECTROGRAM SELECTION FOR SED

We can hear and observe a variety of sounds in terms of how the content changes over time and in frequency. The representation of acoustic time-frequency has many variants, depending on applications and signal-processing methods. Variants may be short-time Fourier transform (STFT) with linear and Mel scales, Constant-Q transforms (CQT), Mel frequency cepstral coefficient (MFCC), or continuous Wavelet transform (CWT). Spectral representations play a crucial role in many applications that use neural networks for classification or regression. One study [25] explains a number of audio filters and their influence at the pre-processing stage of neural network performance and training.

In this sound event detection, a colored log frequency Mel-spectrogram was selected to adopt VODMs owing to its perceptually distinguishable representation. The various audio spectral representations are provided in the supplementary material, Fig. S1. A color map was used to quantify the magnitude of a given frequency within a given time window. The representation was similar to the visual object in an image and the frequency axis was kept constant for audio object detection along the temporal axis. In this way, the VODMs could be applicable to audio representation in terms of domain adoption and fine-tuning. For the 10-second raw audio input, log frequency Mel-spectrogram of size 1000 × 800 with 32 as the number of Mel bins was generated using an 8 kHz sampling rate. Fast Fourier transform (FFT) was applied with a window size of 512 and an overlap of 250 between neighboring windows to extract the spectrogram of audio clips. The sampling rate was only 8 kHz because sounds were recorded at the same rate using CCTV. Details of spectral audio representation are provided in supplementary Section-I. Finally, the generated log Mel spectrogram was inputted to the CNN and VODMs to detect the sound event area along the temporal line.

### III. METHODS

The aim of this study was to develop an autonomous monitoring system for welfare management in large cow farms based on sound onomatopoea. We first applied a conventional CNN structure with certain improvements and then proceeded to two-stage visual object detection for audio by treating acoustic signals as RGB images. This section details the architectural details and evaluation metrics for rare SED using CNN and visual object detectors.

A simple CNN architecture was selected in this experiment that did not involve the use of any CNN variant such as guided learning [2], a recurrent neural network [5], [6], or weakly-supervised learning [21], [22]. These architectures have provided best results in previous challenges on pre-existing data and included annotations. In this experiment, we were dealing with a hard-annotated dataset and wanted to compare the basic CNN modality with visual object detection to investigate whether the novel visual detectors were applicable to R-SED or not. First of all, a CNN structure with minor modifications based on [1] was presented and a comparative study was made with VODMs in two acoustic data environments. An overview of the proposed CNN and applied VODMs is provided in the following sections.
on the fully-connected layer using Softmax nonlinearity and binary cross-entropy loss to compute the probable event area with class level.

### B. VISUAL OBJECT DETECTOR FOR SED

Visual object detection using neural network structures is by now a well-established field and is widely applied in a variety of real-world problem-solving contexts. However, the use of these powerful architectures in audio event detection remains very rare. In this study, we apply a visual object detector for R-SED using a three-channel Mel-spectrogram as input. The application was based on a standard two-stage framework: in the first stage, audio event proposals were generated by a region proposal algorithm, while in the second stage, the proposals were used for audio object classification and localization refinement. A pre-trained ResNet101 [27] structure using ImageNet dataset [28] was first fine-tuned with the audio Mel-spectrogram. Next, the base network feature maps were treated according to different object detection strategies. A short description of these object detection modalities is provided in this subsection, in terms of SED and classification. A schematic overview of the VODMs-based architectures is shown in Fig. 2.

![FIGURE 2. Schematic overview of audio event detector using VODMs.](image)

The FRCNN was an advanced architecture based on Fast-RCNN [29] that integrated a feature extraction network, a region proposal network (RPN), and a classification and regression network within a single network. The feature extraction network was a base network trained on the ImageNet dataset. The RPN was designed to propose multiple sound objects that were identifiable within Mel-spectrogram input. The final classification and regression network was responsible for the detection and classification of sound events. The overall architecture of the system was the same as the original FRCNN structure, the only difference being that the frequency axis was kept constant with sound events only varying on the temporal axis.

The FPN made an improvement to the basic FRCNN architecture by building a scale-invariance pyramid structure on the feature map from the base network. The FPN computed multi-scale versions of the identical Mel-Spectrogram from different feature maps of convolution layers at different spatial resolutions. Sound event prediction gradually increased from low to high-level features of the pyramid. The time axis was only considered for SED and other train-and-test operations were the same as in the original FPN paper.

The CRCNN and cascade FPN (CFPN) were modified versions of FRCNN and FPN, respectively, with sequencing of the detector to reduce noisy bounding box while training by setting multiple increasing thresholds for Intersection over Union (IoU). The final output was the only location in the Mel-spectrogram that exceeded the highest threshold in the sequence.

### C. PERFORMANCE METRICS

Sound event evaluation metrics were based on segments or events. Segment-based metrics compared system output and references at fixed-length intervals, while event-based metrics evaluated them at the event-instance level. Mesaros et al. [30] explains well the common metrics for SED and how they are adapted and interpreted in polyphonic and monophonic cases. The evaluation metric used in our experiments using the CNN sound event detector was the error rate (ER) described in [30], [31] and an $F_1$ score, which is the harmonic mean of precision and recall. VODMs were evaluated using average precision (AP) introduced in the Pascal VOC Challenge [32] and the $F_1$ score.

The ER in equation (1) measures the number of errors based on false positives (FN) and false negatives (FN).

$$ER = \frac{FN + FP}{No. of \ events \ in \ the \ ground \ truth}$$

The $F_1$ score is an overall measure of a model’s accuracy that combines the best precision and recall rate. A good $F_1$ score has low FPs and FNs and is considered perfect when it is 1. Equation (2) calculates the $F_1$ score based on precision and recall.

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

The AP computes the average precision value for recall values from 0 to 1. For a given task and class, the precision-recall curve is computed from a method’s ranked output;
TABLE 3. Evaluation result of CNNs for SED in three test set of cow sound.

| Test data | ER | \(F_1\) | ER | \(F_1\) | ER | \(F_1\) | ER | \(F_1\) | ER | \(F_1\) | ER | \(F_1\) |
|-----------|----|--------|----|--------|----|--------|----|--------|----|--------|----|--------|
| Set-I     |    |        |    |        |    |        |    |        |    |        |    |        |
| CNN_MaxPool_9L [1] | 1.23 | 0.193 | 1.31 | 0.445 | 0.94 | 0.302 | 1.01 | 0.420 | 1.125 | 0.340 |
| CNN_AvgPool_9L [1] | 1.13 | 0.091 | 1.45 | 0.387 | 1.01 | 0.147 | 1.12 | 0.342 | 1.177 | 0.242 |
| Proposed CNN_MaxPool | \textbf{1.01} | \textbf{0.435} | 1.54 | 0.421 | \textbf{0.95} | \textbf{0.311} | \textbf{0.63} | \textbf{0.658} | \textbf{1.031} | \textbf{0.456} |
| Proposed CNN_AvgPool | 1.11 | 0.142 | 1.43 | 0.341 | 1.05 | 0.076 | 0.78 | 0.512 | 1.093 | 0.268 |

# Bold numbers indicate the highest evaluation scores; 9L = 9 Layer CNN structure

TABLE 4. Evaluation result of CNNs for SED in UrbanSound8K test dataset.

| CNNs              | Car horn | Dog bark | Gun shot | Jackhammer | Siren | Mean |
|-------------------|----------|----------|----------|------------|-------|------|
| ER | \(F_1\) | ER | \(F_1\) | ER | \(F_1\) | ER | \(F_1\) | ER | \(F_1\) | ER | \(F_1\) |
| CNN_MaxPool_9L [1] | 1.19 | 0.466 | 1.39 | 0.296 | 0.99 | 0.521 | 0.62 | 0.707 | 0.61 | 0.698 | 0.960 | 0.537 |
| CNN_AvgPool_9L [1] | 1.11 | 0.489 | 1.32 | 0.332 | 0.79 | 0.613 | 0.52 | 0.745 | 0.57 | 0.712 | 0.863 | 0.578 |
| Proposed CNN_MaxPool | \textbf{0.91} | \textbf{0.550} | 1.16 | 0.396 | 0.81 | 0.612 | 0.53 | 0.744 | 0.56 | 0.718 | \textbf{0.795} | \textbf{0.604} |
| Proposed CNN_AvgPool | 1.16 | 0.474 | \textbf{1.13} | \textbf{0.407} | \textbf{0.75} | \textbf{0.635} | \textbf{0.48} | \textbf{0.762} | \textbf{0.54} | \textbf{0.732} | 0.812 | 0.602 |

# Bold numbers indicate the highest evaluation scores; 9L = 9 Layer CNN structure

and average precision is the area under the curve. Recall can be defined as the proportion of all positive examples ranked above a given rank, whereas precision is the proportion of all examples above that rank that are from the positive classes. By varying the confidence threshold, we can modify whether the predicted box is positive or negative. The AP summarizes the shape of the precision/recall curve, and was defined as the mean precision at a set of eleven equally spaced recall levels \([0, 0.1, 0.2, \ldots, 1]\), while mean average precision (mAP) is the AP computed over classes of the sound event. The expression (3) describes the average precision where, \(p_{\text{interp}}(r) = \max_{\tilde{r} \geq r} p(\tilde{r})\) and \(p_{\text{interp}}(r)\) represents the maximum precision for any recall values greater than \(r\). \(p(\tilde{r})\) is the measured precision at recall \(\tilde{r}\) so that the AP is calculated as the mean of \(p_{\text{interp}}(r)\) at all recall levels.

\[
AP = \frac{1}{11} \sum_{r \in [0, 0.1, 0.2, \ldots, 1]} p_{\text{interp}}(r) \tag{3}
\]

IV. EXPERIMENTAL DETAILS

A. CNN BASED COW SED

End-to-end training was performed on the synthetic cow sound dataset with frame-level prediction. Table 3 shows the error rate and \(F_1\) scores between the precision and recall curve in our three test sets of cow sounds. The results show that the proposed CNN with max-pooling performed best in the three test datasets; Set-I (synthetic), Set-II (synthetic plus real), and Set-III (real).

In comparison with previous results [1], the proposed CNN perform better using both max-pooling and average-pooling to reduce the feature map dimensions. For a more robust evaluation of the system, the CNN-based SED was extended for comparison using a new acoustic environment with the UrbanSound8K dataset. The sound events were different from animal sounds both in their frequency and patterns.

B. CNN BASED URBANSOUND8K SED

In this experiment, the same CNN architecture described in Section III-A was used as the sound event detector. Table 4 shows comparative results for CNN detectors using five class categories from the UrbanSound8K dataset. The results show that the proposed CNN performed 3% better than previous methods in terms of \(F_1\) score evaluation.

In the overall system analysis using the cow sound and urban sound datasets, the CNN-based detectors achieved relatively low performance for real-world implementation. We therefore decided to implement a visual detector for cow sound events using time-frequency representation with a color map as an RGB image.

C. COW SED USING VODMS

Visual object detectors were applied for cow sound event detection to evaluate an automated welfare management system based on sound activity. The four detector frameworks were trained on a synthetic cow sound event dataset and
Table 5. Evaluation result of VODMs for SED in three test set of cow sound.

| Test data | VODMs | Bull | Calf | Cow | Mix | Mean |
|----------|-------|------|------|-----|-----|------|
|          |       | AP   |   F₁ | AP  |   F₁ | AP   |   F₁ |
| Set-I    | FRCNN | 0.794 | 0.755 | 0.853 | 0.779 | 0.810 | 0.719 | 0.601 | 0.583 | 0.764 | 0.709 |
|          | FPN   | 0.769 | 0.740 | 0.791 | 0.731 | 0.784 | 0.718 | 0.619 | 0.603 | 0.741 | 0.698 |
|          | CRCNN | 0.830 | 0.776 | 0.824 | 0.777 | 0.819 | 0.752 | 0.714 | 0.672 | 0.797 | 0.744 |
| Set-II   | FRCNN | 0.777 | 0.775 | 0.761 | 0.752 | 0.756 | 0.717 | 0.682 | 0.675 | 0.744 | 0.730 |
|          | FPN   | 0.880 | 0.826 | 0.761 | 0.729 | 0.405 | 0.478 | 0.658 | 0.662 | 0.676 | 0.674 |
|          | CRCNN | 0.856 | 0.821 | 0.772 | 0.741 | 0.489 | 0.544 | 0.656 | 0.628 | 0.693 | 0.684 |
| Set-III  | FRCNN | 0.894 | 0.827 | 0.712 | 0.664 | 0.515 | 0.538 | 0.611 | 0.569 | 0.683 | 0.650 |
|          | FPN   | 0.833 | 0.784 | 0.684 | 0.633 | 0.548 | 0.565 | 0.583 | 0.550 | 0.662 | 0.633 |
|          | CRCNN | 0.888 | 0.847 | 0.718 | 0.684 | 0.567 | 0.561 | 0.633 | 0.622 | 0.702 | 0.678 |
|          | CFPN  | 0.825 | 0.820 | 0.708 | 0.680 | 0.545 | 0.563 | 0.614 | 0.610 | 0.673 | 0.668 |

*Bold number indicate highest evaluation scores.*

The results showed that the CRCNN performed the best on the three test datasets. For Set-I (synthetic), Set-II (synthetic plus real) and Set-III (real), its mAP scores were 0.797, 0.748 and 0.702, respectively; while its $F_1$ scores were 0.744, 0.734 and 0.678, respectively. The IoU with the increasing threshold of CRCNN played a vital role in making robust predictions at a higher stage of the detector. Table 5 shows the mean average precision score at IoU threshold 0.5 and the $F_1$ scores between the precision and recall curve in our three test sets of cow sounds. The CRCNN detector results are visualized in Fig. S2 of the Supplementary section. Comparing the results of cow sound event detection in terms of network structure and CNNs, all of the 2D object detector architectures performed well, with a maximum variation of 7.2% in AP and 6% in the $F_1$ score. For the test set comparison, the performance variation was 6.6%, which was the system mismatch score using real and artificial audio.

The outcome based on the classes, the bull bellowing sound, achieved the highest evaluation score because its repeated sequences appear uniquely in the spectrogram representation. The cow/calf lowing sound was almost similar or indistinguishable (in spectral and wave audio visualization) and even to human ears, with the only difference being in frequency ranges due to the variation in pitch and signal energy. As Fig. S2 of the Supplementary section shows, in certain cases the system results identified the lowing cow sound as calf sounds, so it is clear that the proposed system also appeared to confuse these two class categories. The Mix sound was the sound made of several cattle at the same time, which confused the system and reduced the evaluation scores.

Based on the observed results, an automated welfare management system can be feasibly developed. The system is capable of analyzing sound event frequency and types, and provide information to farm laborers about cattle behavior and actions needed. For example, if the automated system analyzes frequent Mix sounds, this might indicate Hunger or Threat, while if cows or bulls are frequently producing sound then this might indicate Cow in heat or Angry Bull. Onomatopoeic cattle sounds vary according to the animal and frequency range, but identifying the sounds produced by cattle in a restricted environment can be a valuable tool in monitoring their activity and hence in welfare management more generally.

Since it remained unclear whether the proposed system performed well in a wide range of R-SED applications, we extended our study to sound events in urban areas. Section IV-D compares the performance of the CNN-based cow sound event detectors with the Eventness in a new acoustic environment.

### D. URBANSOUND8K SED USING VODMS

The first attempt in SED using FRCNN was done by Eventness using artificial audio dataset with events from the UrbanSound8K dataset. Since the same synthetic dataset used in Eventness was unavailable, we prepared our dataset using the same event classes as for the UrbanSound8K dataset to ensure a fair comparison. The Eventness prepares multi-event synthetic audio by manual selection in pairs of sound events and writes them in temporal position with or without overlapping. Overlapping sound events affect the actual audio information because of the additive amplitude and distorts the accrual harmonics seen in the spectrogram. Therefore, multi-event audio was prepared by maintaining some distance in writing each event randomly along the time axis of the background audio.

Instead of using a single detection model, a comparative analysis of four baseline detector models was made in this experiment to determine the best sound object detector. Again, the CFPN outperformed the others, with an average $F_1$ score of 0.73 and a mean average precision (mAP) of 0.695. Table 6 shows comparative results for the various detectors on our dataset with the Eventness results on its dataset using the same event classes. The $F_1$ scores for Car horn and Gun shot.
using the Eventness achieved the highest evaluation scores even though there were fewer data samples compared to the other classes. On the other hand, the Siren class was unique compared to other classes but the Eventness achieved the lowest evaluation results. Therefore, we concluded that the Eventness system may suffer from overfitting due to its limited number of training samples and diversity in its audio scenes. In our experiment, the same FRCNN detector showed relatively lower performance than the Eventness. The CFPN performed the best for all sound categories, showing uniform results corresponding to the training data scale. Fig. 3 in the Supplementary material section visualizes some detection results using the CFPN in test data samples. In the class-based comparison, our result showed higher performance for Jackhammer and Siren due to their unique acoustic nature, while Car horn and Gun shot suffered from the limited data problem and achieved relatively low evaluation scores.

Overall, the results from the CNN and VODMs detectors showed that performance was dependent on data characteristics and correlation between classes. The results might have been different for other datasets and network architectures but the data preparation and representation according to the problem domain played a vital role using a data-driven approach. The CNN-based SED result showed that the cow sound dataset was unable to adjust to real-world data patterns but using urban sound data the CNN structures found significant (but cannot overcome the VODMs). In the case of the VODMs, the detectors were appropriate for any rare sound event. The system adjusted well to real-world audio patterns using large-scale synthetic data with a little real-world data. In this study, we found that visual object detection architectures performed better than CNN architectures on our two synthetic datasets.

### V. CONCLUSION

This study investigated rare sound event detection through a comparative analysis of conventional CNN with visual object detectors. The experiment was performed using two synthetic datasets for cow sounds and urban sounds. The datasets were rendered more natural by using a wide range of events and scenes, randomly selecting event onset points, and editing multi-event audio so as to avoid events overlapping. In this study, we began by following recent trends in sound event detection by using CNN architectures. Even though our proposed CNN structures outperformed previous results in our two datasets, however, the results were unsatisfactory. This motivated our application of a visual object detection model for sound event detection, using RGB Mel-spectrogram as input. The system evaluation for three cow sound test datasets showed that the applied procedure adjusted well to real-world scenarios, in spite of a limited availability of manually annotated real-world data samples. The variations in prediction result were only 6.6% between the real and synthetic unseen datasets. An experiment using urban sound was conducted to investigate the robustness of the system in a new acoustic environment. Compared to previous studies, the visual object detector performed better and more meaningfully on the real-world unseen samples. In conclusion, visual object detection may be a viable option for rare sound event detection, particularly in locations where the visual image is concealed or invisible or only an audio signal is present.

The application of rare sound event detection to livestock welfare management remains a relatively unexplored topic. Especially for cattle sound event detection, we hope that this study may help to lend momentum to further research in this field. In future, research related to automated analysis and management of cattle may be a promising direction. Similarly, the welfare management system itself could be extendable to other livestock using acoustic information such as pig cough detection or detection alerts of intruders in open-area bird farming.

### ACKNOWLEDGMENT

(Yagya Raj Pandeya and Bhuvan Bhattacharai contributed equally to this work.)

### REFERENCES

[1] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, “Cross-task learning for audio tagging, sound event detection and spatial localization: DCASE 2019 baseline systems,” 2019, arXiv:1904.03476. [Online]. Available: http://arxiv.org/abs/1904.03476

[2] L. Lin, X. Wang, H. Liu, and Y. L. Qian, “Guided Learning convolution system for DCASE 2019 Task 4,” in Proc. DCASE Challenge, 2019, pp. 1–5. [Online]. Available: https://arxiv.org/abs/1909.06178

[3] R. Lu, Z. Duan, and C. Zhang, “Multi-scale recurrent neural network for sound event detection,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 131–135.

[4] S. Adavanne and T. Virtanen, “A report on sound event detection with different bimodal feature,” in Proc. DCASE Challenge, 2017, pp. 1–4. [Online]. Available: http://dcase.community/documents/challenge2017/technical_reports/DCASE2017_Adavanne_130.pdf

[5] W. Ding and L. He, “Adaptive multi-scale detection of acoustic events,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 28, pp. 294–306, 2020.
[6] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, “Convolutional recurrent neural networks for polyphonic sound event detection,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 25, no. 6, pp. 1291–1303, Jun. 2017.

[7] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, “Sound event localization and detection of overlapping sources using convolutional recurrent neural networks,” 2018, arXiv:1807.00129. [Online]. Available: http://arxiv.org/abs/1807.00129

[8] K. Wang, L. Yang, and B. Yang, “Audio events detection and classification using extended R-FCN approach,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 6154–6162.

[9] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The Pascal visual object classes (VOC) challenge,” Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, Jun. 2010.

[11] K. Fauvel, V. Masson, E. Fromont, P. Faverdin, and A. Termier, “Towards sustainable dairy management—A machine learning enhanced method for estrus detection,” in Proc. KDD ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2019, pp. 1–9, doi: 10.1145/3292500.3330712 ffhal-02190790.

[12] M. L. O’connor. (1993). Heat Detection and Timing of Insemination for Cattle. Extension Circular 402. [Online]. Available: https://extension.psu.edu/heat-detection-and-timing-of-insemination-for-cattle