International Journal of Technology 13(6) 1213-1221 (2022)
Received April 2022 / Revised August 2022 / Accepted September 2022

International Journal of Technology

http://ijtech.eng.ui.ac.id

Fusion of Visual and Audio Signals for Wildlife Surveillance

Cheng Hao Ng[1], Tee Connie[1*], Kan Yeep Choo[2], Michael Kah Ong Goh[1]

[1]*Faculty of Information Science & Technology, Multimedia University, 75450, Melaka, Malaysia*
[2]*Faculty of Engineering, Multimedia University, Persiaran Multimedia, 63100, Cyberjaya, Malaysia*

Abstract. Wildlife-vehicle collision (WVC) has been a significant threat to endangered species in Malaysia. Due to excessive development, tropical rainforests and their inhabitants have been edged towards extinction. Road buildings and other linear infrastructures, for instance, have caused forest destruction and forced wild animals to come out from their natural habitats to compete for resources with the human-beings. In Malaysia, much precious wildlife have been lost due to road accidents. Road signs and warning lights have been set up near wildlife crossing, but these do not help much. In this paper, we aim to propose a wildlife surveillance mechanism to detect the existence of wildlife near roadways using visual and audio input. Machine learning classifiers, including Convolution Neural Network (CNN), Support Vector Machine (SVM), K-nearest neighbors (KNN), and Naive Bayes, are adopted in the study. We focus on five types of the most frequently occurring wildlife on the roads: elephants, tapirs, Malayan bears, tigers, and wild boars. Experimental results demonstrate that a good accuracy as high as 99% can be achieved using the proposed approach. On the other hand, the Naïve Bayes classifier ranks the lowest in performance with an accuracy value only up to 86%.

Keywords: Deep learning; Fusion; Machine learning; Wildlife surveillance

1. Introduction

Studies show that Malaysia has lost many precious widlife species due to road accidents (Jantan et al., 2020). For example, around 102 tapir had been killed by road accidents in the last decade. Because of the total tapir population of only about 1,000 left in Peninsular Malaysia, such fatal accidents impose significant losses. The problem of wildlife-human conflict is worsening because the forests have been cleared to make way for development like road infrastructure building. Rapid growth has adversely altered wildlife profiles and destroyed their natural habitats. The extensive invasion of the wildlife ecosystem has pushed them to extinction.

In response to the rise of wildlife-vehicle collision (WVC) in the country, the Wildlife and National Parks Department has set up road signs at wildlife crossing areas. Besides, solar lights and transverse bars have also been installed at the crash-prone zones. Nevertheless, these measures are sometimes ineffective due to the natural responses of the wild animals when encountering human-being or vehicles. Instead of fleeing, wild animals would sometimes become immobilized and experience inescapable shocks when caught in a traumatic situation (Zanette et al., 2019). It might be too late to escape when they run

*Corresponding author's email: tee.connie@mmu.edu.my, Tel.: +606-2523592, Fax.: +606-2318840
doi: 10.14716/ijtech.v13i6.5876

into vehicles. Drivers' honking or flashing lights would not chase them away but instead causes them to freeze and be hit by the cars.

To address WVC, this paper presents a fusion of visual- and audio-based wildlife surveillance approach using machine learning methods. Automatic monitoring is performed at wildlife frequently occurring areas to detect potential wild animals that will go near the roadway. Instead of passively letting the wilderness enter the highway which might cause possible WVC, the proposed approach takes proactive measure to warn the related department to chase the animals away before they come near the roadway. In this study, we focus on five animal species: elephant, tapir, tiger, Malaysian bear, and wild boar. These animals are commonly reported to be observed near man areas.

Surveillance systems that rely on visual input, e.g., CCTV cameras, are easily affected by illumination and pose changes. For example, the appearance of the animals' changes when viewed from different angles (e.g., front, back, left, and right). Besides, the scene also changes when acquired during different times of the day, , e.g., the scene appears clear under bright sunlight, but the scene becomes invisible during the night. Therefore, there is a need to complement visual-based recognition with audio-based input to improve the reliability of the proposed system. In this study, a pipeline approach is presented to process both the visual and audio input for the animals using AI and machine learning techniques (Berawi, 2020; Siswanto, 2022; Fagbola, 2019). Discriminative features are extracted from both types of signals and the extracted features are fused using deep learning approach. Experimental results show that the fusion of visual and audio signal can greatly improve the overall accuracy of the proposed method.

2. Methods

2.1. Size of Dataset

In this study, we focus on five types of animals: tapir, elephant, Malayan bear, tiger, and wild boar. The dataset are mainly collected from Internet sources. It is difficult to find sources that contain both image and audio sound. Therefore, the pictures and audio files are downloaded separately from the Internet. The image data sources mainly come from Kaggle and Google Images. As for sound sources, it is tough to collect, and it is almost not available for certain animals such as tapir and Malayan bear. As a result, the videos of the animals are downloaded from YouTube. After that conversion, the video files were converted manually to audio files. After that, the audio signals were segmented to derive multiple samples from a single audio file. In the end, there are 215 samples each for both image and audio signals, for each type of animals. So, there are a total of 1075 (215 x 5) samples each for image and audio data. Some sample images are illustrated in Figure 1.

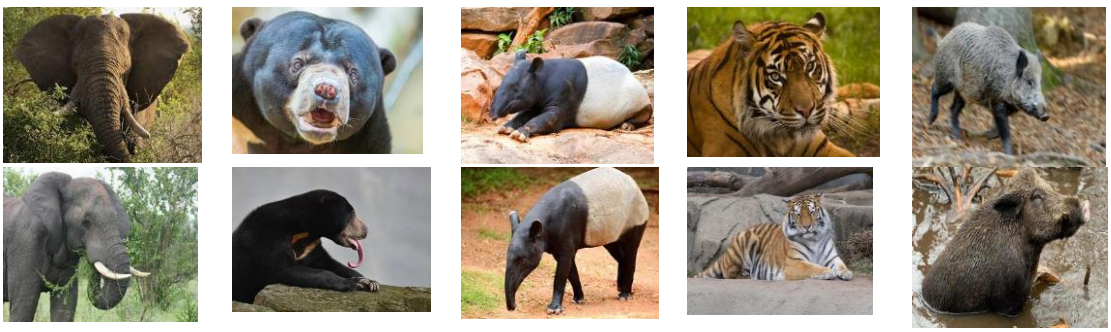

Figure 1 Sample images in the dataset. Left column to right: elephants, Malayan bear, tapir, tiger, wild boar

2.2. *Pre-processing*

The collected audio files are trimmed so that all the files' duration is standardized to one second. The audio files are converted to WAV file format, with a sampling rate of 16000 HZ, mono channels, and 16-bit depth. To expand the dataset for audio signals, data augmentation is further performed. Various techniques like addition of distribution noise, shift time, stretching, change of speed, and change of pitch are applied.

To pre-process the image files, the region of interest containing the animals are manually cropped from the image. The purpose of performing this step is to remove unnecessary background from the pictures. In addition, augmentation is also applied to increase the image samples. After that, all the images are converted to size 150 x 150 pixels.

2.3. *Audio Signal Processing*

The audio signals are represented using Mel spectrogram features (Ulutas et al., 2022). Mel spectrogram is one of the most potent features used to learn the time and frequency representation from the audio sequences. Several processes are involved in calculating Mel spectrogram and Fast Fourier Transform (FFT) is one the important processes. Given N number of audio samples, the general expression to calculate the output of FFT is given as (Equation 1),

$$F_k = \sum_{n=1}^{N} f(n) \cos\left(\frac{2\pi nkT}{N}\right) - i \sum_{n=1}^{N} f(n) \sin\left(\frac{2\pi nkT}{N}\right) \quad (1)$$

where $f(n)$ refers to the n-th sample value, k is the discrete frequency variable, and T denotes the signal period.

The input to Mel spectrogram is WAV files. The output spectrogram illustrates the signal's intensity (or "loudness") at various frequencies in a waveform across time. Figure 2 and Figure 3 show some sample audio signals and the spectrograms for the different species of animals, respectively.

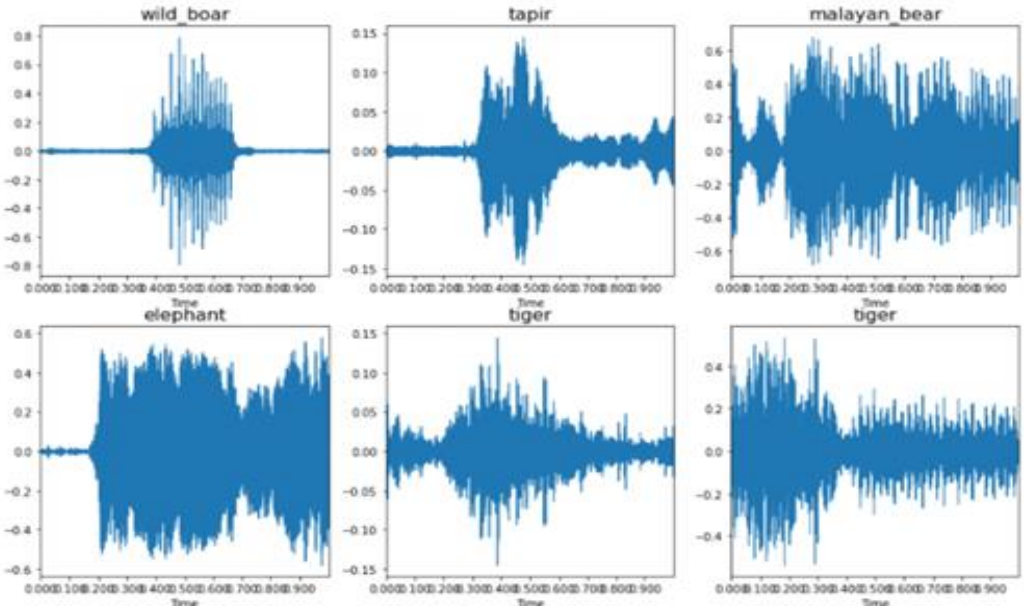

Figure 2 Samples of original audio signals

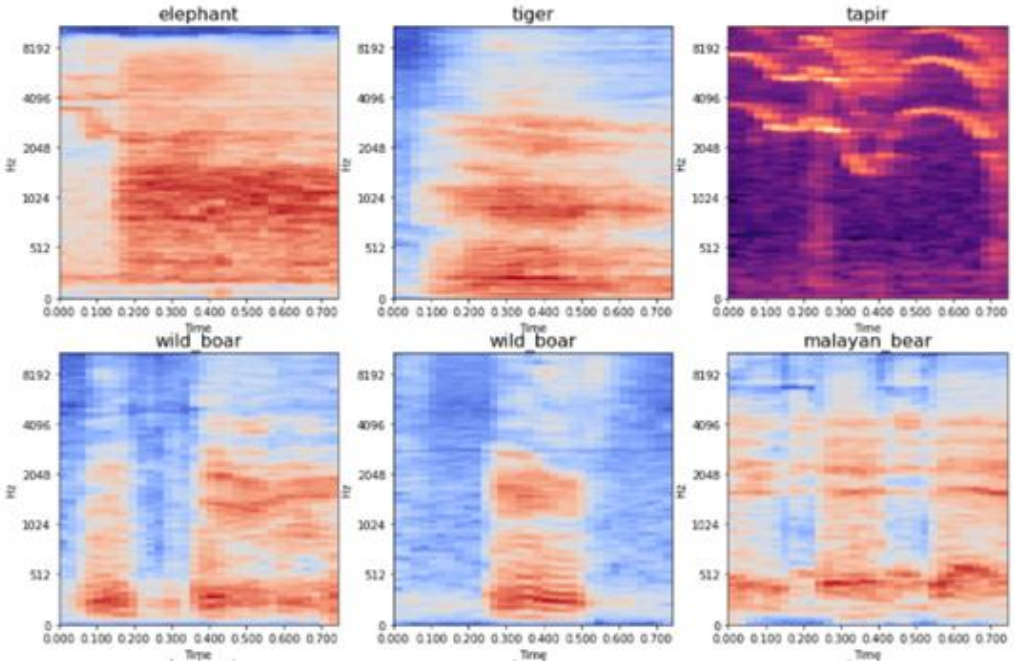

Figure 3 Samples of spectrograms corresponding to the audio signals

Data augmentation is applied on the audio signals to broaden the dataset size further to synthesize additional samples. The techniques used for data augmentation include addition of distribution noise, shift time, stretching, change of speed, and change of pitch. Sample audio signals, and their corresponding spectrogram after data augmentation are displayed in Figure 4.

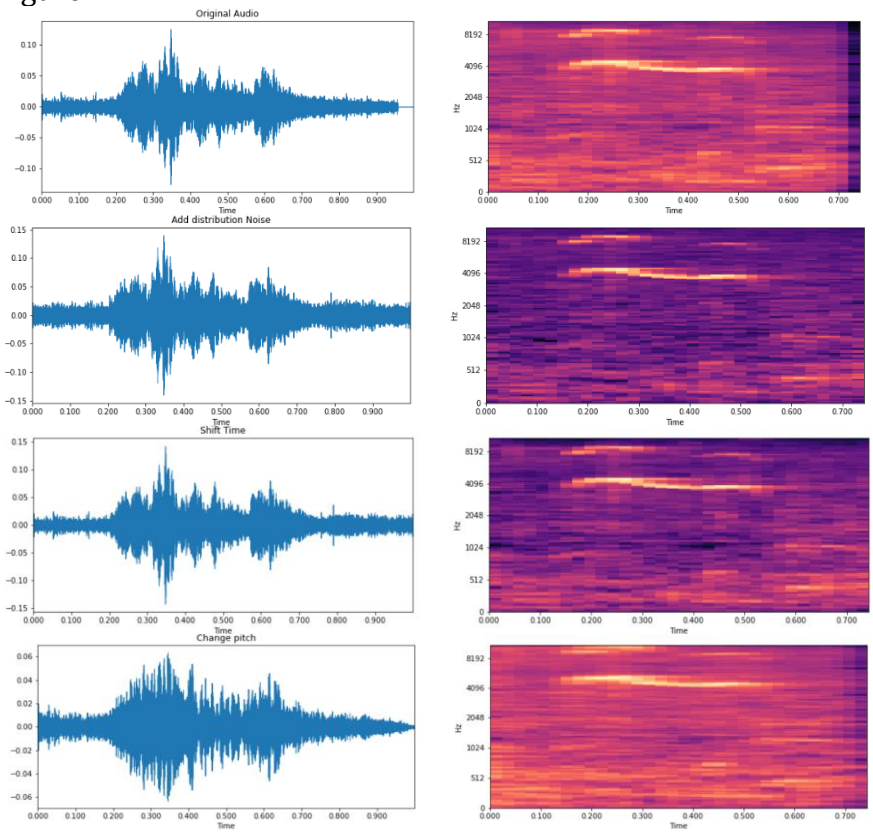

Figure 4 Samples of audio signals (left) and the corresponding spectrogram (right) after data augmentation. From top to bottom: Original image, after noise addition, stretching, change of pitch

2.4. Visual Image Processing

To process the visual images, this project adopts Principal Component Analysis (PCA) (Fromentèze et al., 2022) to extract discriminative features from the wildlife images. PCA is a well-known dimensionality reduction technique that transforms high-dimensional data into a reduced space that maintains the multitude of values of the more extensive set.

Given a set of N training images of the wild animals, $X_i = [x_1, x_2, \dots, x_N]$, the covariance matrix, C_x, is first determined. After that, a linear transformation V is found to transform the original dataset X_i into a new subspace Y. The generalized eigenvalue problem is given as (Equation 2),

$$C_x = VDV^T \tag{2}$$

where $D = diag(\lambda_1, \lambda_2, \dots, \lambda_N)$ is the diagonal matrix and $U = (u_1, u_2, \dots, u_N)$ are the eigenvectors. Then, the PCA transformation can be computed as (Equation 3),

$$Y_i = V^T X_i \tag{3}$$

where $i = 1,2, \dots, M$ and M denotes the total number of pixels in the image.

2.5. Fusion and Classification of Audio- and Visual-based Features

After the audio and visual samples have been processed, they are fused together to perform classification. In this paper, feature-level fusion is adopted. In feature-level fusion, feature sets from the two input sources are combined into a single feature set. The real benefit of feature-level fusion is that it is able to encode prominent characteristics that might increase classification performance by detecting associated feature values provided by the distinct input modalities.

Firstly, all input images are resized to the same width and high of 150 × 150 pixels. Then PCA is applied to extract unique features from the input images. On the other hand, the audio samples are converted from waveform to spectrogram. A spectrogram can be viewed as a pictorial representation of the audio signals representing a spectrum of frequencies of the movement. Therefore, the spectrogram is also processed by PCA to obtain a compact audio signal representation. The last step is to concatenate the feature of image and audio together becomes a new feature.

The extracted visual and audio features are concatenated to form an extended vector, $Z_i \in \mathbb{R}^{d \times f}$ where d and f are the dimension of the transformed output features using PCA for the visual and audio signals, respectively. After that, a classifier is used to classify Z_i into one of the animal classes. In this study, many classifiers, including Convolutional Neural Networks (CNN) (Gautam & Singhai, 2022), Support Vector Machines (SVM) (Ansari et al., 2021), Naïve Bayes (Khamdamovich & Elshod, 2021), and K-nearest neighbours (KNN) (Saleem & Kovari, 2022) are evaluated. Figure 5 shows the overall process in the proposed method.

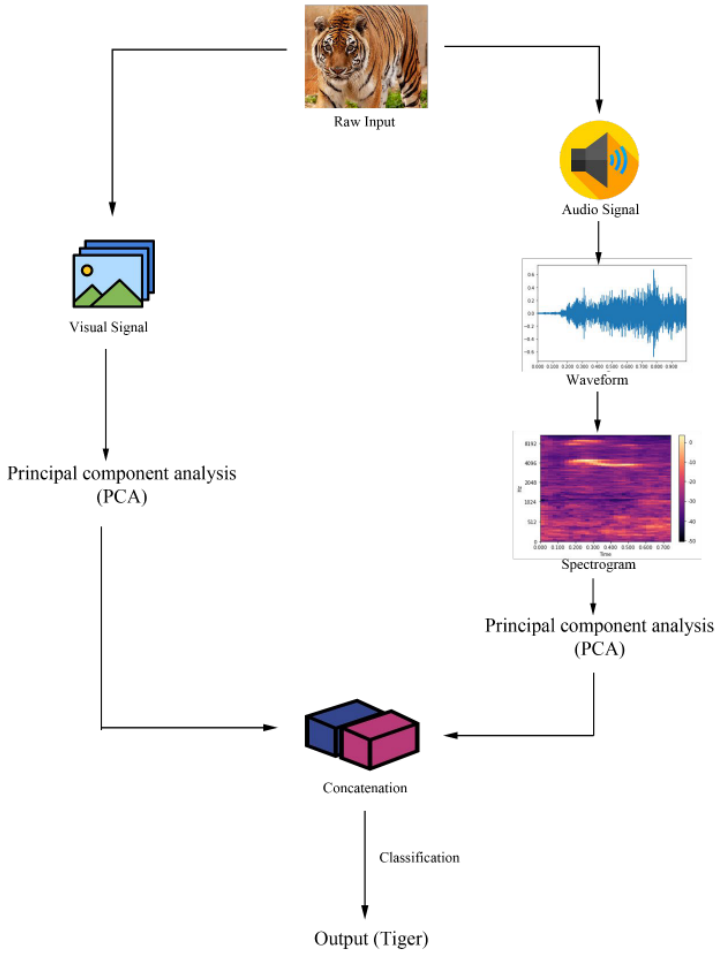

Figure 5 Overall processes involved in the audio- and visual-based classification system

3. Results and Discussion

3.1. Experiments for Audio Signals

The first set of experiments are conducted to evaluate the audio signals. The audio signals are separated into training and testing data using an 80%-20% split. The audio samples are first converted to waveform and subsequently to spectrogram representation. After that, PCA is applied to obtain the feature representation for the audio signals. The PCA feature is then fed to the classifiers to perform prediction. Four classifiers are tested, namely CNN, SVM, Naïve Bayes and KNN.

In this paper, the deployed CNN model is configured as depicted in Table 1. The ReLU activation function is used for the first three layers. The Adam optimizer is adopted and the Sparse Categorical Crossentropy loss function is utilised. Test accuracy of 99% and a loss of 0.0196 are reported using the proposed architecture. The training and validation accuracies and training and validation loss are provided in Figure 6.

In SVM, the RBF kernel is utilised with a regularization parameter of 1. As for the Naïve Bayes classifier, the value of the smoothing variance is set to $1e^{-9}$. Next, for k-NN, Euclidean distance is used with the value for k set to 3. A comparison of the different experiments using the audio signals is provided in Table 2. Overall, CNN and SVM yield the best results. Naïve Bayes classifier, on the other hand, performs the worst. This might be due to its inability to work with the non-linear boundaries of the audio data.

Table 1 Architecture of CNN used in the experiments

Layer (type)	Output Shape	Number of Parameters
Conv1d (ConvID)	(79, 128)	384
Conv1D_1 (ConvID)	(78, 64)	16 448
Conv1d_2 (ConvID)	(77, 32)	4 128
Max_pooling1d (MaxPooling1D)	(38, 32)	0
Flatten (Flatten)	(1, 1 216)	0
Dense (Dense)	(1, 128)	155 776
Dropout (Dropout)	(1, 128)	0
Dense_1 (Dense)	(1, 64)	8 256
Dropout_1 (Dropout)	(1, 64)	0
Dense_2 (Dense)	(1, 5)	325

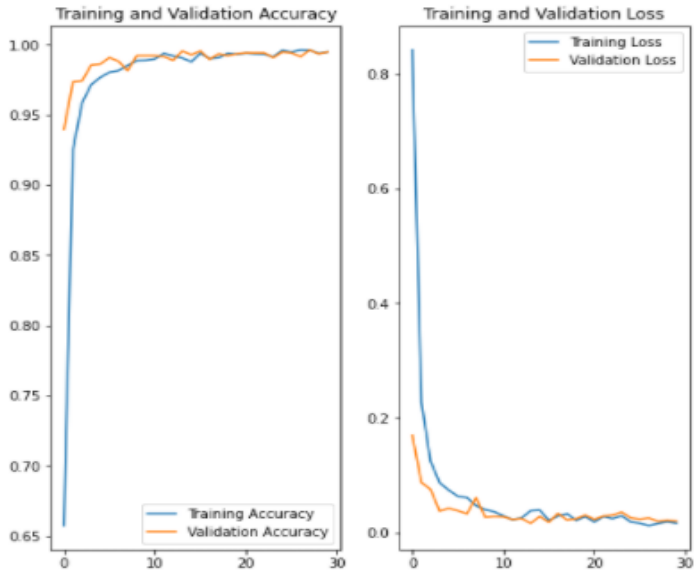

Figure 6 Training and validation accuracy (left) and training and validation loss (right) of using CNN

Table 2 Number of receptors in each container

Model	Average Accuracy (%)
CNN	99.47
SVM	99.73
Naïve Bayes	83.39
KNN	98.60

3.2. Experiments for Visual Signals

The experiments for visual samples are conducted using almost the same settings as described in the experiments for the audio signals, except for CNN. In the visual-based experiments, LeakyReLU activation function with the alpha 0.1 is used instead of the ReLU function. The other settings remain the same. The settings for SVM, Naïve Bayes, and KNN are the same as the audio-based experiments.

The experimental results for the visual-based samples are presented in Table 3. We observe that SVM has an accuracy of 81.26%, and it is the highest among all. This is because SVM works well with a clear margin of separation and is effective in high-dimensional spaces. Overall, image-based samples' performance is much inferior to audio-based signals. This is due to the large appearance variations in the visual data. For example, the images of

an elephant look very different when viewed from the front and the back. Moreover, the image dataset is confounded by changes in illumination.

Table 3 Comparison for visual samples

Model	Average Accuracy (%)
CNN	80.27
SVM	81.26
Naïve Bayes	60.66
KNN	77.08

3.3. Experiments for Fusion of Audio and Visual Signals

In this section, the performance of the proposed fusion approach is assessed. For CNN implementation, the architecture illustrated in Table 1 is used. Again, the same settings for SVM, Naïve Bayes and KNN and adopted. The performance of fusing the audio and visual signals using the different classifiers is shown in Table 4. We observe that SVM had the highest accuracy before the fusion method compared to other classifiers. After applying the fusion method, the CNN model achieves the highest accuracy. We speculated the reason is that the fused feature has indeed provided valuable information to complement the recognition accuracy. The Naïve Bayes classifier again ranks the lowest in performance.

Table 4 Comparison of fusion approaches

Model	Average Accuracy (%)
CNN	99.27
SVM	98.74
Naïve Bayes	86.98
KNN	98.34

3.4. Discussions

There are several exciting findings in this study. Firstly, we find that The audio-based samples yield good results (above 80%) for all the different classifiers. The reason for this is mainly due to the fact that the audio dataset used contains clear animal sounds without much background noises. This makes it easier to perform sound recognition.

On the other hand, the performance using the image dataset is lower than the audio dataset. This is because the images of the animals contain different confounding factors (like pose and illumination changes) that affect the appearances of the animals.

In the image-based experiments, elephants' and wild boars' images are often confused with each other. This is most probably because the appearance of an elephant and wild boar are similar in size and color, especially when viewed from a distance.

Feature fusion has greatly enhanced the performance of image-based recognition. The CNN model can achieve an accuracy of up to 99%. The fused feature has provided valuable information to complement the recognition accuracy of the image-based approach.

4. Conclusions

A wildlife animal surveillance system using the fusion of audio and visual signals are presented in this paper. The proposed system can recognize the wild animal and warn the road user when an animal is detected near the road. An accuracy of 99% can be achieved when the audio and visual signals are fused. In the future work, more comprehensive fusion approaches will be explored, including data level fusion and score level fusion schemes. Besides, other deep learning approaches will also be investigated to achieve better classification results.

Acknowledgements

This project is supported by the TM R&D Fund, 2022 (Grant no. MMUE/220023).

References

Ansari, M.R., Tumpa, S.A., Raya, J.A.F., Murshed, M.N., 2021. Comparison Between Support Vector Machine and Random Forest for Audio Classification. *In*: 2021 International Conference on Electronics, Communications and Information Technology (ICECIT), pp. 1–4

Berawi, M.A., 2020. Managing Artificial Intelligence Technology for Added Value. *International Journal of Technology,* Volume 11(1), pp. 1–4

Fagbola, T.M., Thakur, C.S., Olugbara, O., 2019. News Article Classification using Kolmogorov Complexity Distance Measure and Artificial Neural Network. *International Journal of Technology*, Volume 10(4), pp. 710–720

Fromentèze, T., Yurduseven, O., Del-Hougne, P., Smith, D.R., 2022. Principal Component Analysis for Microwave and Millimeter Wave Computational Imaging. *In*: 2022 International Workshop on Antenna Technology (IWAT), pp. 72–75

Gautam, S., Singhai, J., 2022. Fast and Accurate Water Region Classification from Remote Sensing Images Using Enhanced Convolutional Neural Network Classifier. *In*: 2022 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). pp. 1–4

Jantan, I.D., Ghani, Y., Makhtar, A.K., Isa, M.H.M., Jawi, Z.M., 2020. A Study on Animal and Vehicle Collisions in Malaysia Based on News Analysis. International Journal of Road Safety, Volume 1(2), pp. 63–69

Khamdamovich, K.R., Elshod, H., 2021. Detecting Spam Messages Using the Naive Bayes Algorithm of Basic Machine Learning. *In*: 2021 International Conference on Information Science and Communications Technologies (ICISCT), pp. 1–3

Saleem, M., Kovari, B., 2022. A Comparison Between K-Nearest Neighbor and Jk-Nearest Neighbor Algorithms for Signature Verification. *In*: 2022 21st International Symposium INFOTEH-JAHORINA (INFOTEH), pp. 1–5

Siswanto, J., Suakanto, S., Andriani, M., Hardiyanti, M., Kusumasari, T.F., 2022. Interview Bot Development with Natural Language Processing and Machine Learning. *International Journal of Technology*, Volume 13(2), pp. 274–285

Ulutas, G., Tahaoglu, G., Ustubioglu, B., 2022. Forge Audio Detection Using Keypoint Features on Mel Spectrograms. *In*: 2022 45th International Conference on Telecommunications and Signal Processing (TSP), pp. 413–416

Zanette, L.Y., Hobbs, E.C., Witterick, L.E., MacDougall-Shackleton, S.A., Clinchy, M., 2019. Predator-Induced Fear Causes PTSD-Like Changes in the Brains and Behaviour of Wild Animals. *Scientific Reports,* Volume 9(1), pp. 1–10