Outdoor Long Range Ubiquitous Projectiles Tracking System Using P-MPLR and Computer Vision

Minjae Kim
Kyung Hee University
Yongin, Republic of Korea
kmj5596@khu.ac.kr

Hansu Jeong
Sangmyung University
Seoul, Republic of Korea
201710982@sangmyung.kr

Jeongwon Moon
Kyungpook National University
Daegu, Republic of Korea
bella7365@knu.ac.kr

Donghyeon Na
Sangmyung University
Seoul, Republic of Korea
201721402@sangmyung.kr

Woojin Choi
Sun Moon University
Asan, Republic of Korea
twinsno119@sunmoon.ac.kr

Sophia Lee
Purdue University
West Lafayette, United States
lee3155@purdue.edu

Ethan O'Sullivan
Purdue University
West Lafayette, United States
eosulli@purdue.edu

Anthony Smith
Purdue University
West Lafayette, United States
ahsmith@purdue.edu

***Abstract*—A growing body of research has been dedicated to investigating systems for quickly detecting and pinpointing the location of projectiles aimed at a target. Various studies and products have demonstrated the feasibility of predicting the point of origin for a shot. However, existing tools come with limitations and inconveniences. This study aims to introduce a model that achieves highly accurate and low-latency tracking of shot origins during long-range shooting, free from constraints, by incorporating LoRa, computer vision, and sound detection technologies. The proposed system commences with an initial photograph captured by a camera positioned near the target. Whenever the adjacent acoustic sensor detects a fired projectile, the camera captures a new image, triggering the Raspberry Pi located beside the target to calculate the coordinates of the shot by comparing it with the previous image. When compared to the actual bullet mark, the predicted bullet mark's error remains under 1 mm. This paper's proposal showcases accuracy suitable for integration into military intelligence training systems, offering a cost-effective solution.**

***Index Terms*—LoRa, Computer vision, Sound detection**

I. Introduction

In 2017, the Small Arms Survey found that the United States had 120.5 firearms per 100 residents, surpassing the population itself [1]. The National Rifle Association hosts over 11,000 shooting tournaments and 50 national championships annually, catering to all ages. Target shooting's popularity extends to the military, where precision matters. Despite dedicated military training, target verification remains tedious, involving manual marking and comparison. Proposed research aims to streamline this process and holds promise in addressing the issue.

ShotMarker constitutes a software application tailored for the F-class shooting championship, an event conducted exclusively in the prone position spanning distances from 275 m to 915 m [2]. Nonetheless, this system is not devoid of complexities. Firstly, external factors such as wind significantly impact its precision. Although the typical error under ideal conditions rests between 2 to 3 mm, this rate of inaccuracy can increase when unfavorable weather prevails during shooting. Secondly, the system is confined to supersonic bullets. In essence, if a firearm employs subsonic ammunition, the functionality of ShotMarker might falter. This paper introduces a new approach to the realm of projectile mark detection systems. The proposed system attains impressive accuracy while circumventing a multitude of environmental constraints. Notably, unlike ShotMarker, which is restricted to subsonic projectiles, this system breaks free from bullet type limitations. It accomplishes this by deducing shooting coordinates through a comparison of images before and after firing. This distinctive attribute grants versatility in accommodating various projectile types within the system. Similar to ShotMarker, the approach incorporates Long Range (LoRa) network technology to facilitate long-distance shots. The proposed methodology involves capturing an initial image to serve as a reference, followed by the transmission and superimposition of coordinates onto this reference image. Despite slower transmission pace of LoRa, our system prioritizes coordinate data over image data, mitigating the impact on transmission speed and diminishing packet loss concerns.

In outdoor tests, potential packet loss due to environmental factors is acknowledged. To ensure reliable packet delivery, the Multi-Packet LoRa (MPLR) algorithm is used, countering such losses. MPLR enhances LoRa's appeal due to extended battery life, long-range communication, and cost-effective equipment [3]. The system's distinctive feature is its reset after each bullet shot, facilitating swift bullet grouping calculation. Demonstrating precision, it's well-suited for military intelligence training systems.

II. Relative Works and Motivation

The suggested study comprises three primary technologies: LoRa, computer vision, and sound detection.

Andreea developed an innovative electronic archery scoring system that involves calculating the variance between successive images using dual cameras [4]. Zin et al. took this approach further by not only employing background subtraction but also utilizing frame differences between the current frame and the two immediate frames to identify new arrows [5]. In contrast to [4], a critical aspect to address is the optimal selection of thresholds for image segmentation. They introduced dynamic thresholding after image binarization, utilizing a total of two threshold values. The integration of color-based background modeling, gray-level frame differences, and dynamic thresholding collectively contributes to precise target and arrow detection.

The effectiveness of Park et al.'s approach in accurately processing outdoor images is hindered by environmental conditions like brightness [6]. This research centers on utilizing YOLO V5 to concentrate on the target, detecting a gunshot based on target alterations. Despite the persistent focus on the target by the camera, challenges persist in detecting gunshots amidst diverse backgrounds.

Li et al. discuss the improvement of low illumination images [7]. In their work, the RGB color space was transformed into YUV color space to extract illuminance and reflection elements. They further isolated the illuminance and reflection components using homomorphic filtering to preserve image edges and minimize noise. Global adaptive gamma correction was then employed to adjust image brightness, resulting in the enhancement of low illumination images. This study employed a similar approach, enhancing outdoor images through gamma correction to extract the target.

As shown in [8], there is a case of sound classification using the Convolutional Neural Networks (CNN) method by converting sound into a spectrogram. Similarly, our proposed research employs a moderately shallow CNN model, minimizing training demands. In a previous experiment, a CNN achieved an 86.7% accuracy in classifying 10 categories using the UrbanSound8K dataset. This involved 3 fully connected (FC) layers, impacting parameters and computational complexity. In contrast, our research streamlines calculations by reducing FC layers and focuses on binary classification.

Chen et al. introduced an MPLR approach that effectively recovers from packet loss during LoRa-based image transmission [9]. The time taken for image transmission in LoRa is influenced by factors like spreading factor (SF), bandwidth (BW), coding rate (CR), and image size. Faster transmission occurs with lower SF, higher BW, CR of 1, and smaller image sizes. For instance, employing MPLR under specific conditions such as SF 7, BW 500, CR 4/5, and an image size of 9 KB led to a transmission time of 4.92 seconds without packet loss. Even under the same parameters, when 2% packet loss occurs, MPLR achieves swift recovery in just 0.17 seconds, resulting in a total recovery time of 5.09 seconds.

The proposed paper addresses limitations by advocating for a novel system encompassing three key innovations: (1) extended-range outdoor capability, (2) minimized latency and packet loss, and (3) compatibility with a wide range of projectiles. Leveraging LoRa's long-range communication advantages, this system effectively manages shadow effects, enhancing precise image analysis, representing novelty (1). The P-MPLR protocol not only reduces latency but also safeguards against packet loss in LoRa communication. This becomes especially crucial for outdoor long-range communication, where interference from various frequencies is prevalent. Novelty (3) surpasses prior research by eliminating constraints on projectile types, offering a versatile solution.

III. Methodology

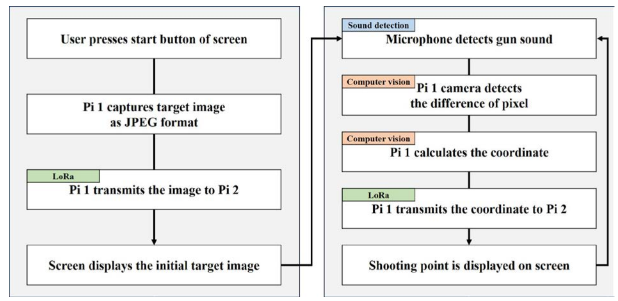

Fig. 1. Illustration of the proposed method.

The illustration of the proposed method is portrayed in Fig. 1. The proposed system consists of 2 Raspberry Pi 4B 8GB, 2 sx1262 LoRa Hat for Raspberry Pi, Raspberry Pi HQ Camera Module and HP-DK40 Microphone. The system process is divided into two steps: initial stage and continuous stage. In the initial stage, Pi 1 camera captures the target image and sends it to Pi 2. In the continuous stage, the shooting point is decided based on the difference of pixel between the previous and the after image on occasion a gunshot sound is detected.

A. LoRa

LoRa uses less power and limited bandwidth, therefore, it displays a low transmission speed to send a large portion of data such as image or video. Despite this shortage, compared to Wi-Fi and Zigbee, LoRa is suitable for data transmitting over long distances [10]. Image transmission is mandatory for the proposed system considering initial target visualization. To reduce LoRa transmission time, this paper focused on decreasing the image size in an initial stage and the shooting point coordinate is sent instead of an image in the continuous stage. Transmitting coordinates requires 40 times fewer data units than transmitting images. Also, LoRa communication itself has a disadvantage of packet loss because no protocol has been applied. There is no guarantee that data is safely transmitted to the receiver because it communicates only on the physical layer. For stable image transmission, thus, a protocol capable of ensuring packet transmission is required. Therefore, the proposed system applied the P-MPLR protocol, which eliminated the unnecessary portion of point-to-point communication in the MPLR protocol.

B. Computer Vision

This paper aims at fast transmission using LoRa communication while reducing the effect on light after target image capture. To execute the goal, the background portion is removed through a series of processes. Shooting point is determined by the comparison of the previous image. Based on the pixel subtraction, an image negative transformation and gaussian blurring are performed. Then, image emphasis is made through divide operation. Finally, the shooting point can be obtained using the contour. Thereafter, only the point is transmitted to the shooter. The image of the target is taken in the same environment for each shooting along with the image being updated and stored for each new repetition.

C. Sound Detection

The system acknowledges a gunshot when a gunshot sound is detected, afterwards, figure out the coordinate of the shooting point. Mel-spectrogram is used for sound detection. By utilizing Short Time Fourier Transform, a sound wave that has time amplitude domain is converted to a spectrogram that additionally has a frequency domain. This is because the characteristic of the audio data is shown in the frequency domain. Mel-spectrogram is deduced by applying Mel-filter bank to spectrogram. This converts the frequency value lower than 1,000 Hz as linear, and the frequency value higher than 1,000 Hz as log-scale. Mel-spectrogram image is used for learning CNN model, and binary classification is done that determines whether the sound is gunshot sound or not.

The CNN that detects gun sounds consists of neurons that self-optimize through learning. The proposed paper focuses on advancing the classification performance of the gun sound by reducing the parameters required to train with fewer hidden layers to prevent overfitting. After the convolution operation, FC layer classifies whether the gun-sound. This paper proposes the adoption of the Exponential Linear Unit (ELU) that improves the gradient vanishing problem of the activation function [11]. Batch normalization (BN) was applied instead of dropout to reduce the effect of covariate shift, where the distribution of the activation function output changes according to the change of the weight parameter in the previous layer [12]. Finally, the model is applied with an adaptive gradient algorithm for efficient optimization [13]. It has advantage of not having to manually adjust learning rate.

D. Data Augmentation

The most crucial problem in sound analyzing research is the lack of data training. It causes overfitting and difficulty in handling unseen data [14]. Therefore, in this research spectrogram augmentation such as time warping, time masking and frequency masking prevents the overfitting of the model and improves the performance. The augmented data with SpecAugment makes the model solid and reduces the effects of partial loss in the frequency and time domain information.

IV. Implementation

A. Dataset

The dataset consisted of audio data obtained from multiple sources [15], [16]. Also, included are shotgun sounds recorded directly from the shooting range. To generate a standardized spectrogram, the dataset was organized in units of 2 seconds. A categorical analysis of the data is reported in TABLE I. Gunshot sounds have a proportion total of 43%. The dataset is divided into 3 categories: training data, validation data, and test data; 64%, 16%, and 20%, respectively.

TABLE I
AUDIO SAMPLES CATEGORICAL LISTS

Class	# of data
Gun Shot	6,560
Air Conditioner	1,000
Car Horn	429
Children Playing	1,000
Dog Bark	1,000
Drilling	1,000
Engine Idling	1,000
Jackhammer	1,000
Siren	929
Speech	581
Street Music	1,000

Additionally, frequency masking was used to augment the spectrogram data in this research. This function augments the data by concealing part of the frequency domain which is the y-axis of the spectrogram. As for the concealed range, the frequency masking parameter value was set to 80 as used in [14]. Frequency masking was applied to all spectrogram images, and the number of data was increased from 15,499 to 30,998. Fig. 2 depicts what a power spectrogram looks like after being augmented.

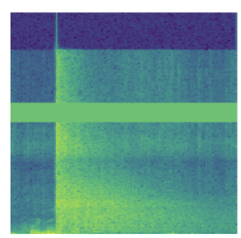

Fig. 2. Power-spectrogram after applying augmentation.

B. Gunshot sound classification

The system utilizes the torchaudio Python library for audio analysis. Each audio differs from the sample rate and quantization level; the channel is mixed as mono and stereo with various audio lengths. Thus, the dataset was unified as a mono channel, 44,100 Hz sampling rate and 2 seconds duration using zero padding and slice.

A lot of information can be obtained when the sound data is analyzed from a frequency point of view. The Time-Magnitude domain of the sound data is converted to the frequency domain.

Each sound data source of spectrogram was overlapped by 50% in a window unit of 23.2 ms to constitute 87 frames and 1,024 Fast Fourier Transform (FFT) points were used for each frame. It is converted into a 223 $\times$ 217 spectrogram using transform module. Additionally, in the case of the Mel-spectrogram, in accordance with the calculation of 128 Mel band energy per frame, frames are converted to 223 $\times$ 217 size of Mel-spectrogram [17]. Apply the log scale on the amplitude squared of the extracted spectrogram and Mel-spectrogram and convert it to decibel units. Fig. 3 illustrates the spectrogram and Mel-spectrogram features from one data.

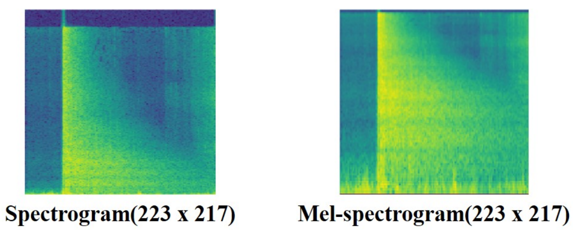

Fig. 3. An example of feature extraction.

This paper applies a CNN model for gunshot sound classification using the PyTorch framework. The architecture of the CNN network is described in detail in TABLE II. Network is contained of 4 convolutional blocks and a FC layer. Each convolutional block consists: BN, ELU activation function, and maxpooling with 2 $\times$ 2 kernel size and stride value as 1. Model training is proceeded with cross entropy loss function, 50 epochs, 0.001 initial learning rate and 256 batch size for model parameter update. The model accuracy is 99.1%, recall is 98.29%, precision is 99.58% and F1-Score is 98.93%.

TABLE II
DESCRIPTION OF THE CNN MODEL ARCHITECTURE

Layer	Output Shape	Channel Size	Param #
Input Layer	217 $\times$ 223	3	0
Conv2D	218 $\times$ 224	8	104
BatchNorm2D	218 $\times$ 224	8	16
MaxPool2D	109 $\times$ 112	8	0
Conv2D	110 $\times$ 113	16	528
BatchNorm2D	110 $\times$ 113	16	32
MaxPool2D	55 $\times$ 56	16	0
Conv2D	56 $\times$ 57	32	2,080
BatchNorm2D	56 $\times$ 57	32	64
MaxPool2D	28 $\times$ 28	32	0
Conv2D	29 $\times$ 29	16	2,064
BatchNorm2D	29 $\times$ 29	16	32
MaxPool2D	14 $\times$ 14	16	0
Linear	2	-	6,274

C. Networking

LoRa is a modulation technique that operates solely upon the physical layer of the open systems interconnection model. This means that implementing LoRa on its own offers no ability for clients to coordinate transmissions to avoid interfering with each other, and delivery of data also cannot be guaranteed. Solutions to these two issues were built into the proposed system.

The issue of interference avoidance was solved by limiting communications to a single point to point link operating within a single channel in the 915 MHz frequency space. Also, frequency interference could be solved by using the MPLR protocol. The proposed system is only designed for point-to-point communication that does not require unnecessary overhead of MPLR. Due to the limited packet size of LoRa, the proposed protocol needs to minimize header size. The structure of the packet is shown in Fig. 4.

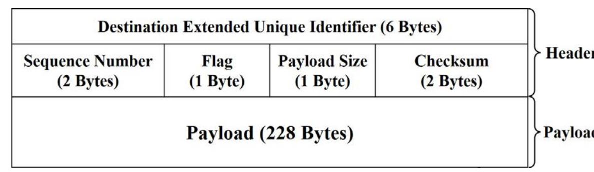

Fig. 4. Structure of the proposed packet protocol.

Packet has 228 bytes of payload size and 12 bytes of header size. The destination EUI has a MAC address of the destination. Mac address is a unique address which every device has. If multiple shooters are using this proposed system very close to each other, there needs to be assurance that transmissions are unique to each other. Therefore, the proposed protocol has 6 bytes destination EUI unlike 4 bytes destination EUI of MPLR. The sequence number is used to check the order of the packets. The flag indicates packet type. In the flag, there are SYN, SYN-ACK, DATA, BVACK, ACK, FIN flags [9]. The payload size is the length of payload. The checksum is existing to check integrity for each packet.

D. Computer Vision

Image is generated by a combination of light reflected from the surface of the object or light reaching the camera. According to a mathematical expression of the illumination-reflection model, each pixel value F(x,y) of the image is made by the multiplication of the illumination component I(x,y) and the reflection component R(x,y). While the illumination component is mainly concentrated in low frequency band, the reflection component is concentrated in high frequency band. In order to separate these two components, homomorphic filtering is performed as follows. First, converting RGB to YUV, log operation is taken for the Y component to generate Y'. By applying FFT to Y', the illumination and the reflection component can be controlled in the frequency domain, respectively. Secondly, Y' is multiplied by low pass and high pass filter, to separate Low Frequency (LF) and High Frequency (HF) components. Next, a scaling factor is applied to LF and HF component to adjust the illuminance and reflection. An image in the form of a spatial domain is generated through Inverse FFT and exponential operation is performed in the adjusted image. YUV is changed as a result of applying homomorphic filtering to Y' [7].

TABLE III
IMAGE TRANSMISSION SPEED AND PACKET LOSS RATE ACCORDING TO THE APPLICATION OF THE PROTOCOL AND IMAGE PROCESSING

Location	Protocol	Image Processing	Temperature	Humidity	Wind	Distance	Transmission Time	Packet Loss
Urban	No	No	2 °C	48%	3 km/h	300 m	11 min	30%
Urban	No	No	8 °C	44%	7 km/h	850 m	11 min	90%
Urban	No	No	26 °C	44%	10 km/h	1 km	11 min	15%
Rural	No	Yes	12 °C	64%	10 km/h	300 m	70 sec	10%
Urban	No	Yes	3 °C	80%	3 km/h	300 m	70 sec	30%
Rural	Yes	Yes	5 °C	90%	3 km/h	300 m	80 sec	No
Urban	Yes	Yes	4 °C	72%	18 km/h	300 m	100 sec	No

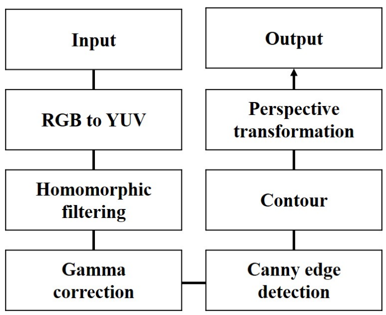

Fig. 5. Image processing diagram to obtain initial target portion.

Subsequently, gamma correction is used to achieve the constant intensity normalization effect regardless of brightness. And, canny edge detection algorithm is applied to obtain clear target edge. Finally, a target portion can be acquired by performing perspective transformation to project a 3D scene image into 2D. The overall process is illustrated in Fig. 5. Also, Fig. 6 shows the captured image and the result of detecting the target.

Fig. 6. Original image and the processing result of the image.

E. Application

Pi 2 performs byte binding on image packets transmitted from server of the Pi 1. The original image taken by the camera, only parts of target are imported, an image smaller than the original is sent to the client server. In the process of importing the target image, the quality of the truncated target image is downsampled to improve packet transmission speed. Moreover, for user visibility, image quality is improved on the web. In the proposed system, when the camera detects a bullet mark, it detects the center coordinate of the bullet mark and transmitted to a web server. It draws a bullet mark on the target by multiplying the target's size, which is described in the image conversion, by the received coordinate values. To find a shot group, the distance between the two bullets must be found by comparing the distances of all the bullets stored by the shooter. Prior to that, since the unit used in the present system is a pixel, unit conversion was performed in inches through scale conversion. The length provided in pixels was changed to inches before the group size was measured, in which case the Pythagorean theorem was used.

F. Experiments

The experiment was divided into two parts: protocol performance measurement and system performance evaluation. The protocol performance measurement focused on the image processing time and packet loss, depending on the application of P-MPLR for acquiring initial images before shooting. The system performance evaluation involved applying P-MPLR to recognize firearm sounds during shooting at a distance of 100 meters and accurately detecting the position of the bullets.

1) Protocol Performance Measurement: This experiment was performed to compare the transmission speed and packet loss rate of the initial target image based on the application of the protocol before shooting. It was conducted at two distinct locations: an urban setting (Purdue University) and a rural area (13626 S 525 W, Romney, IN 47981). Constraints of experiment include whether to apply P-MPLR, whether to apply image processing and weather conditions. The image processing was carried out following the procedure outlined in Fig. 5. The original image size was about 120 KB, but after image processing, the image size was reduced to about 11 KB. The detailed experimental procedure and results are described in TABLE III. Without applying the protocol and image processing, there was a maximum packet loss of 90%, and image transmission was not possible regardless of the distance. Under the same conditions, at a distance of 5 m indoors, there was no packet loss, but it still took 11 minutes to transmit the image. When only image processing was applied, an average packet loss of 20% occurred. After applying both P-MPLR and image processing, it took an average of 90 seconds to transmit the image without any packet loss.

2) System Performance Evaluation: This experiment was conducted exclusively at the farm three times. The firing distance was 100 m, using 5.56 mm caliber bullets. The angle of the camera is 30 degrees above the ground, and a model trained under the conditions outlined in Table II was employed to detect gunshots. Each shooting session involved firing 10 rounds, and the recognition accuracy was 100%.

The initial target image extracted after starting the system was 11 KB, and it took 80 seconds. When the gun shot, it took about 5 seconds to determine that it was a gunshot. After determining that it was a gunshot, it took about 2 seconds for the projectile position to be displayed on the web. Fig. 7 represents the predicted projectile displayed on the web and the actual projectile. Furthermore, the distance error between the shooting location and the coordinates transmitted through image processing was within 1 mm.

Fig. 7. Predicted projectile and the actual projectile.

V. Conclusion

This research suggests the feasibility of long-distance communication in diverse conditions and the ability to verify shooting outcomes through a persistent tracking system. Through the integration of LoRa technology and the implementation of P-MPLR, a communication framework devoid of packet loss was established. Leveraging computer vision technology, the system effectively extracted target images with background removal, facilitating accurate determination of the shooting location. The system's dependability was substantiated via a series of experiments conducted across different environments.

Nonetheless, this study presents certain limitations. Despite the application of homomorphic filtering, parameter configuration remains challenging due to the dynamic variations in outdoor illuminance and reflection levels, which can fluctuate based on different scenarios. Moreover, image transmission through LoRa proves to be time-inefficient when compared to alternative communication methods.

To enhance the system's performance, there is a requirement to enhance the reliability of target extraction and conduct additional experiments that closely simulate real-world scenarios. Furthermore, this research proposes the logging of each shooter's performance onto a server database, making it applicable for applications such as military training or shooting competitions.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW(2018-0-01865, 2019-0-01880 and 2023-0-00042) supervised by the IITP(Institute of Information & Communications Technology Planning & Evaluation) and the Korean MSIT (Ministry of Science and ICT), under the National Program for Excellence in SW)(2021-0-01082) supervised by the IITP(Institute of Information communications Technology Planning Evaluation).

References

[1] A. Karp, "Estimating Global Civilian-held Firearms Numbers," Small Arms Survey, 2018 [Online]. Available: https://www.smallarmssurvey.org/resource/estimating-global-civilian-held- firearms-numbers

[2] A. MacDonald, ShotMarker Manual. (2020). [Online]. Available: https://autotrickler.com/pages/support

[3] A. S. Rawat, J. Rajendran, H. Ramiah and A. Rana, "LoRa (Long Range) and LoRaWAN Technology for IoT Applications in COVID-19 Pandemic," *2020 International Conference on Advances in Computing, Communication & Materials (ICACCM)*, 2020, pp. 419-422, doi: 10.1109/ICACCM50413.2020.9213067.

[4] D. Andreea, "OnTaget: An Electronic Archery Scoring," 2021, *arXiv preprint arXiv:2104.01622*.

[5] T. T. Zin, I. Oka, T. Sasayama, S. Ata, H. Watanabe and H. Sasano, "Image Processing Approach to Automatic Scoring System for Archery Targets," *2013 Ninth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2013, pp. 259-262, doi: 10.1109/IIH-MSP.2013.73.

[6] S. Park et al., "Feasibility of Measuring Shot Group Using LoRa Technology and YOLO V5," *2022 IEEE Sensors Applications Symposium (SAS)*, 2022, pp. 1-6, doi: 10.1109/SAS54819.2022.9881356.

[7] S. Li, H. Guo, W. Sun and X. Sun, "A low-illuminance image enhancement method in YUV color space," *2022 14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 2022, pp. 286-291, doi: 10.1109/ICMTMA54903.2022.00062.

[8] F. Demir, D. A. Abdullah and A. Sengur, "A New Deep CNN Model for Environmental Sound Classification," *IEEE Access*, vol. 8, pp. 66529-66537, 2020, doi: 10.1109/ACCESS.2020.2984903.

[9] T. Chen, D. Eager and D. Makaroff, "Efficient Image Transmission Using LoRa Technology In Agricultural Monitoring IoT Systems," *2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, 2019, pp. 937-944, doi: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00166.

[10] S. Devalal and A. Karthikeyan, "LoRa Technology - An Overview," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, pp. 284-290, doi: 10.1109/ICECA.2018.8474715.

[11] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*. [Online]. Available: https://arxiv.org/abs/1511.07289

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv:1502.03167*, 2015, [Online]. Available: http://arxiv.org/abs/1502.03167.

[13] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, no. 7, pp. 1–39, 2011.

[14] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," *in Proc. Annu. Conf. Int. Speech Commun. Assoc.*, pp. 2613-2617, 2019.

[15] J. Salamon, C. Jacoby and J. P. Bello, "A Dataset and Taxonomy for Urban Sound Research," Nov. 2014. [Online]. Available: https://urbansounddataset.weebly.com/urbansound8k.html

[16] *Gunshot Audio Forensics Dataset*, Development of Computational Methods for the Audio Analysis of Gunshots, NIJ grant 2016-DN-BX-0183, Jun. 2018. [Online]. Available: http://cadreforensics.com/audio/

[17] Z. Chi, Y. Li and C. Chen, "Deep convolutional neural network combined with concatenated spectrogram for environmental sound classification," *2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, pp. 251-254, 2019.