Influence of Video Dynamics on EEG-based Single-Trial Video Target Surveillance System

Heon-Gyu Kwak

Dept. of Artificial Intelligence Korea University Seoul, Republic of Korea hg_kwak@korea.ac.kr

Sung-Jin Kim

Dept. of Artificial Intelligence
Korea University
Seoul, Republic of Korea
s j kim@korea.ac.kr

Hyeon-Taek Han

Dept. of Artificial Intelligence Korea University Seoul, Republic of Korea ht han@korea.ac.kr

Ji-Hoon Jeong

Dept. of Computer Science Chungbuk National University Cheongju, Republic of Korea jh.jeong@chungbuk.ac.kr Seong-Whan Lee
Dept. of Artificial Intelligence
Korea University
Seoul, Republic of Korea

sw.lee@korea.ac.kr

Abstract—Target detection models are one of the widely used deep learning-based applications for reducing human efforts on video surveillance and patrol. However, the application of conventional computer vision-based target detection models in military usage can result in limited performance, due to the lack of sample data of hostile targets. In this paper, we present the possibility of the electroencephalography-based video target detection model, which could be applied as a supportive module of the military video surveillance system. The proposed framework and detection model showed prospective performance achieving a mean macro F_β of 0.6522 with asynchronous real-time data from five subjects, in a certain video stimulus, but not on some video stimuli. By analyzing the results of experiments using each video stimulus, we present the factors that would affect the performance of electroencephalography-based video target detection models.

Keywords-electroencephalography, single-trial event-related potentials detection, video target surveillance, asynchronous brain-computer interface systems.

I. INTRODUCTION

Over the past few years, deep learning has been improved in revolutionary advances across various domains, with video recognition and analysis standing out as one of its most promising applications. Specifically, video target detection plays an important role in urban safety surveillance, traffic management, and various commercial applications [1]–[3]. However, in specialized fields such as military surveillance and patrol, conventional computer vision (CV)-based detection methodologies exhibit limitations due to the limited number of hostile target samples [4].

To address this challenge approaches leveraging human perception and cognition using brain computer-interface (BCI)

This work was supported by the Agency For Defense Development Grant Funded by the Korean Government (UI233002TD).

techniques on the target detection task have been proposed recently [4], [5], as the effectiveness of BCI techniques has been proved [6]–[8]. Among various neuroimaging methods, electroencephalogram (EEG) would be a reasonable method for implementing the target detection system using brain signals, due to its advantages in mobility, low cost, and high time-resolution [9], as many studies presented in various fields of real-time BCI applications [10]–[17].

In this paper, we propose the framework for implementing a video target detection model based on EEG, which can possibly be the supportive module to cooperate with CV based target detection system. Furthermore, by analyzing the trained model, we present insights in constructing the framework for EEG-based video target detection.

II. MATERIALS AND METHODS

A. Subjects

A total of 13 healthy male and female subjects participated in the EEG data acquiring task. All subjects were aged between 20 and 40 years old and had no neurophysiological or mental disorders. Four of the subjects participated in the offline EEG data acquiring process, which is used for training dataset of the target detection model. The other nine subjects participated in the online testing experiments which are for evaluating the model performance in a real-time asynchronous environment.

B. Video Target Stimuli

Three video clips 'Video1', 'Video2-N', and 'Video2-AI' are used as stimuli for acquiring EEG signals from subjects. Each video clip is simulated video surveillance footage, assumed to be recorded in a mountain forest area. Each of these clips is divided into quarters, which shows various scenes

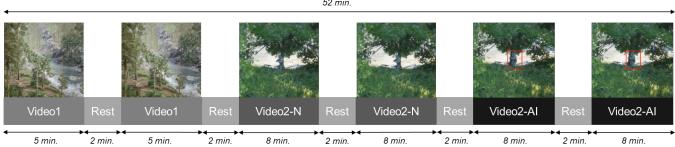


Fig. 1. An overview of the EEG data acquisition process. A subject first watches 'Video1' twice, and then watches 'Video2-N' and 'Video2-AI' twice in turn. Each video clip is eight minutes long. The subject takes rests for two minutes after a session ends. Each subject proceeds four sessions of EEG acquisition, taking 52 minutes in total.

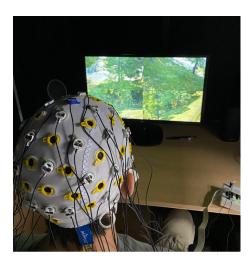


Fig. 2. An example of an EEG data acquiring session with surveillance video clips. Subjects were instructed to concentrate on the presented video clip to detect appearing targets. EEG signals were recorded using 32-channel electrodes, in the soundproof experimental booth with lights off.

TABLE I DIFFERENCES IN VIDEO STIMULI

Dynamics	Name		
	Video1	Video2-N	Video2-AI
Camera rotation	X	0	О
Weather changing	X	0	О
Target bounding box	X	х	О

of surveillance cameras in different settings of angles and positions. 'Video1' is 300 seconds long, while 'Video2-N' and 'Video2-AI' are 480 seconds long.

For Video2-N' and 'Video2-AI', surveillance cameras rotate for three seconds to monitor other sites every five seconds periodically, in turn. Moreover, the weather of the surveillance location changes into foggy or rainy as time goes by.

In each video clip, three types of targets 'deer', 'wild boar', and 'soldier' suddenly appear and last for one second several times, in a specific time point and location. Deer and wild boars are non-hostile targets which can be classified as errortargets, and soldiers are hostile targets representing true-target.

Two video clips 'Video2-N' and 'Video2-AI' are essentially identical except for one difference, 'Video2-AI' shows a red bounding box that guides the locations of targets when they appear, whereas 'Video2-N' does not. 'Video1' has different timing, location, and the number of times in target appearance with 'Video2-N' and 'Video2-AI'. Table I shows the differences between video clips.

C. EEG Data Acquisition Process

Each of the four subjects was instructed to concentrate on video clips to detect each target while equipping the EEG acquiring devices. EEG signals were recorded at a 250 Hz of sampling rate, with the BrainVision Recorder software provided by BrainProduct (GmbH, Germany), using 32 Ag/AgCL electrodes placed in 10-20 international systems. During the acquisition, the impedance of each electrode was maintained under 10 k Ω for excluding noises caused by poor contactivity of electrodes. To minimize artifacts, the recording proceeded in a soundproof experimental booth while the lights were off, and the subjects were asked to minimize facial and body movements. Each subject watched all of the video clips twice each, six sessions of watching in total, and had two minutes of resting time when a session ended. Fig. 1 represents the overview of the configuration of the EEG data acquisition process, and Fig. 2 shows the example of an EEG acquiring session.

D. Model Training

We hypothesized that there would be general EEG patterns related to the target detection across subjects. Whereas, biosignals like EEG have high variations in patterns between individuals [18]-[21], even if they perform the same task or are in the same condition, resulting from unique neurological and physiological characteristics of individuals. To consider both of these factors, we pre-trained the model using all of the subject's EEG data, and fine-tuned model parameters with calibration when testing.

1) Model Architecture: We modified the DeepConvNet [22] into hierarchical architecture, which classifies non-target and targets (true and error) as binary classes first, and re-classifies true- and error-target after it. The main reason for two-stage classification is that we hypothesized that classifying nontarget and target is a relatively easy task compared to classifying true-target and error-target, thus, separating learning parameters would result in more stable performance.

2) Training Objective: We defined the target detection problem of this study as a highly imbalanced 3-class classification task, with the classes of 0 (non-target), 1 (true-target), and 2 (error-target). Therefore, we used multi-class cross-entropy loss for training the model. The multi-class cross-entropy loss is defined as follows:

$$\mathcal{L}(\hat{y}, y) = -\sum_{k=0}^{2} y^c \log \hat{y}^c, \tag{1}$$

where y is the true label of the input data, \hat{y} is the predicted label of the input data, and c is the index of classes of the input dataset.

- 3) Target Label Assigning: For the data of an EEG acquisition session, the total number of data time points is sampling rate times video length. For example, the EEG data acquired with 'Video1' has 75,000 time points ($T=[t_1,t_2,t_3,\ldots,t_{74999},t_{75000}]$). In this set of time points T, a target c appears on a certain time point t_x and lasts for one second. Thus, the label y^c would be assigned to the set of time points $[t_x,t_{x+1},\ldots,t_{x+249},t_{x+250}]$, therefore, the set of labels $Y^c_{t_x}=[y^c_x,y^c_{x+1},\ldots,y^c_{x+249},y^c_{x+250}]$, where $y^c_x=[t_x,y^c]$. For any other time points that are not assigned by class 1 or 2, we assigned 0 as a label.
- 4) Data Augmentation: Compared to the total duration of the video clip, the duration of appearances for each target is enormously small. Even each type of class appears 30 times each on the 'Video2-N', the class ratio between three classes is non-target (87.5 %), true-target (6.25 %), and error-target (6.25 %). This kind of highly imbalanced dataset would drive the model to learn the biased features of the majority class and lower the generalized performance of the model. To alleviate this problem, we augmented data of the minority class using the sliding window method which is known as effective data augmentation method for EEG data [23]. We used the window size with 250 time points with a stride of 25 time points, to overlap the input data in 0.1 seconds of interval. As a result, the volume of the minority data is augmented to ten times larger than its original volume.
- 5) Training Hyperparameters: We optimized the trainable parameters of the model using Adam optimizer [24]. Hyperparameters used for the training were batch size of 128, learning rate of 0.001, weight decay of 0.0001, dropout rate of 0.1, and 100 training epochs.

E. Evaluation Metric

To consider the imbalance of each class, The performances of the model were evaluated with macro F_{β} , which can be defined as

$$recall = \frac{true \ positive}{true \ positive + false \ negative}, \tag{2}$$

$$precision = \frac{true\ positive}{true\ positive + false\ positive},$$

$$(3)$$

$$F_{\beta} = (1 + \beta^2) \cdot \frac{recall \cdot precision}{(\beta^2 \cdot recall) + precision}, \tag{4}$$

macro
$$F_{\beta} = \frac{1}{N} \sum_{c=0}^{N} F_{\beta}^{c},$$
 (5)

where N is the number of classes, F_{β}^{c} is the F_{β} of the class c, and β is the parameter to adjust weights of recall and precision. In this study, N is three because we defined the problem as a 3-class classification task, and we set β as two to give larger weight on the recall.

III. EXPERIMENTAL SETTINGS

We conducted experiments to evaluate the performance of the proposed EEG-based video target detection model in the online environment setting. Nine subjects who did not engage in the offline EEG data acquiring process participated in the online test session. To receive EEG signals asynchronously in real-time, we leveraged the remote data access function of the BrainVision. Out of nine subjects, five subjects watched 'Video1' as video stimuli, and the rest of the four subjects used 'Video2-N' and 'Video2-AI' as video stimuli.

IV. RESULTS AND DISCUSSIONS

A. Online Inference Performance

In the experiments with 'Video1' for five subjects, the proposed model achieved the macro F_{β} of 0.7000, 0.5238, 0.7568, 0.5238, and 0.7568 for subject 5, 6, 7, 8, and 9 respectively, and the overall performance was 0.6522. When the model was trained with the EEG data of 'Video2-N', the performance resulted in 0.1200, 0.1569, 0.0392, and 0.1967 for subject 10, 11, 12, and 13 respectively. The model trained with the EEG data from 'Video2-AI' showed the performance of 0.0784, 0.1091, 0.1200, 0.0952, and 0.1007 for subject 10, 11, 12, and 13. The model trained with EEG data of 'Video1' showed the highest performance, while the data of 'Video2-AI' resulted in the lowest performance. Table II shows the overall performances of the experiments.

The model trained with the EEG data of 'Video1' showed promising performance in the EEG-based target detection task. In contrast, the model trained with the EEG data of 'Video2-N' and 'Video2-AI' showed poor performance, even if other factors that affect model performance, such as the model architecture, the number of parameters, and training hyperparameters, were set in the same condition with the model of 'Video1'. For this performance gap, we analyzed the model and input EEG signals to identify any other factors that might affected the results.

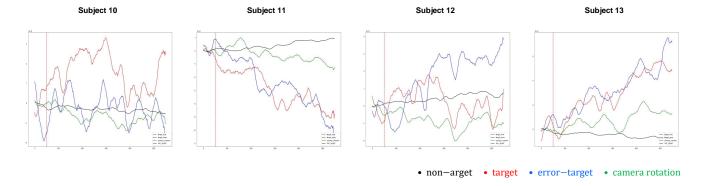


Fig. 3. Grand average of ERPs for subject 10, 11, 12, and 13 by class, with EEG data of 'Video2-N'. The figure contains the ERPs for three seconds of duration after the target appearance. The red vertical line indicates the time point of the target appearance. The black plot is for the ERPs of non-target, the red plot is for true-targets (enemy soldier), blue plot is for error-targets, and green plot is for camera rotations

TABLE II
ONLINE INFERENCE PERFORMANCE BY VIDEO STIMULUS

Performance (macro F_{β})				
Subject	Video Stimulus			
Subject	Video1	Video2-N	Video2-AI	
S5	0.7000	-	-	
S6	0.5238	-	-	
S7	0.7568	-	-	
S8	0.5238	-	-	
S9	0.7568	-	-	
S10	-	0.1200	0.0784	
S11	-	0.1569	0.1091	
S12	-	0.0392	0.1200	
S13	-	0.1967	0.0952	
Mean	0.6522	0.1282	0.1007	

S is the abbreviation of subject.

B. Event-related Potentials (ERPs) Analysis

ERPs are commonly considered important EEG feature that highly influences the decision of the EEG-based target detection models [4], [5]. Therefore, we calculated the grand average of all trials of ERPs from 'Cz', 'C3', and 'C4' channels, to identify if the ERPs played an important feature of the model. Fig. 3 represents the result of the ERPs by classes, in camera rotation. In the figure, subject 11 and 13 showed better performances compared to the subject 10 and 12, but their ERPs showed less clearly distinct differences in patterns between classes. This result implies that the model did not count temporal patters of ERPs as important features, for classifying each class.

C. Saliency Map Analysis

We also analyzed the model using a saliency map [25], to visualize the importance of each EEG channel for the model inference. In Fig. 4, the area colored in deep blue represents that the model performance decreases when the EEG channel data of the area is removed, meaning the EEG channel plays an important role for the trained model in inference. The

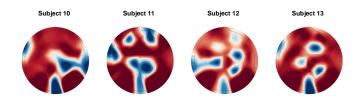


Fig. 4. Channel-wise saliency map of subject 10, 11, 12, and 13 with EEG data of 'Video2-AI'. Deep blue colored areas indicate that EEG channels located in the area are relatively more important than other channels in model inference. Red colored areas indicates vice versa

EEG channels of the red area represent the relatively lower importance of the channel compared to others. As shown in the figure, important channels for the trained model were mainly are located in central, temporal and occipital regions, which are related to the passive visual perception, like camera rotation, rather than the frontal and temporal-parietal regions which are mostly related to dissciminative response task, such as, target detection [26], [27].

These results would be caused by differences in video dynamics, such as camera rotation and weather changes, arousing the passive visual perception-related EEG features more than the discriminative response task-related EEG features to the subject [28]. Imbalanced inductions of EEG features would lead the model to be biased on passive visual perception-related EEG features, therefore, the performance of target detection would be poor.

V. CONCLUSION

In this paper, we presented the possibility of an EEG-based video target detection system in real time asynchronous environment, while showing the system would fail when video stimuli have specific characteristics, like repeated periodically video dynamics that causes passive visual perception, leading the model to train with irrelative EEG features for target detection. To implement a more stable and precise EEG-based target detection system, researchers should carefully design the

experimental paradigm and protocol. We hope that this study will give insights to researchers, and help to advance the field of EEG-based target detection studies.

REFERENCES

- A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis et al., "Deep learning for computer vision: A brief review," Comput. Intell. Neurosci., vol. 2018, 2018.
- [2] G. Sreenu and S. Durai, "Intelligent video surveillance: A review through deep learning techniques for crowd analysis," *J. Big Data*, vol. 6, no. 1, pp. 1–27, 2019.
- [3] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A survey of deep learning and its applications: A new paradigm to machine learning," *Arch. Comput. Methods Eng.*, vol. 27, pp. 1071–1092, 2020.
- [4] Y. Zhang, H. Zhang, X. Gao, S. Zhang, and C. Yang, "UAV target detection for IoT via enhancing ERP component by brain computer interface system," *IEEE Internet Things J.*, 2023.
- [5] X. Song, B. Yan, L. Tong, J. Shu, and Y. Zeng, "Asynchronous video target detection based on single-trial EEG signals," *IEEE Trans. Neural* Syst. Rehabil. Eng., vol. 28, no. 9, pp. 1931–1943, 2020.
- [6] J. Kim et al., "Abstract representations of associated emotions in the human brain," J. Neurosci., vol. 35, no. 14, pp. 5655–5663, 2015.
- [7] K.-H. Thung et al., "Conversion and time-to-conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion," Med. Image Anal., vol. 45, pp. 68–82, 2018.
- [8] A. Supratak, H. Dong, C. Wu, and Y. Guo, "DeepSleepNet: A model for automatic sleep stage scoring based on raw single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 1998–2008, 2017
- [9] M. Lee, C.-B. Song, G.-H. Shin, and S.-W. Lee, "Possible effect of binaural beat combined with autonomous sensory meridian response for inducing sleep," *Front. Hum. Neurosci.*, vol. 13, pp. 425–440, 2019.
- [10] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of EEG signals and facial expressions for continuous emotion detection," *IEEE Trans. Affect. Comput.*, vol. 7, no. 1, pp. 17–28, 2015.
- [11] S.-H. Lee, M. Lee, J.-H. Jeong, and S.-W. Lee, "Towards an EEG-based intuitive BCI communication system using imagined speech and visual imagery," in *Conf. Proc. IEEE Int. Conf. Syst. Man Cybern. (SMC)*, 2019, pp. 4409–4414.
- [12] S.-H. Lee, M. Lee, and S.-W. Lee, "Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 12, pp. 2647–2659, 2020.
- [13] M. Perslev et al., "U-Sleep: Resilient high-frequency sleep staging," NPJ Digit. Med., vol. 4, no. 1, p. 72, 2021.
- [14] S.-B. Lee et al., "Comparative analysis of features extracted from EEG spatial, spectral and temporal domains for binary and multiclass motor imagery classification," *Inf. Sci.*, vol. 502, pp. 190–200, 2019.
- [15] L. S. Vidyaratne and K. M. Iftekharuddin, "Real-time epileptic seizure detection using EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 11, pp. 2146–2156, 2017.
- [16] R. Mane et al., "FBCNet: A multi-view convolutional neural network for brain-computer interface," arXiv preprint arXiv:2104.01233, 2021.
- [17] J.-S. Bang, M.-H. Lee, S. Fazli, C. Guan, and S.-W. Lee, "Spatio-spectral feature representation for motor imagery classification using convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 7, pp. 3038–3049, 2021.
- [18] F.-C. Lin, L.-W. Ko, C.-H. Chuang, T.-P. Su, and C.-T. Lin, "Generalized EEG-based drowsiness prediction system by using a self-organizing neural fuzzy system," *IEEE Trans. Circuits Syst. I: Regul. Pap.*, vol. 59, no. 9, pp. 2044–2055, 2012.
- [19] H.-I. Suk, S. Fazli, J. Mehnert, K.-R. Müller, and S.-W. Lee, "Predicting BCI subject performance using probabilistic spatio-temporal filters," *PLoS One*, vol. 9, no. 2, p. e87056, 2014.
- [20] X. Li et al., "Exploring EEG features in cross-subject emotion recognition," Front. Neurosci., vol. 12, p. 162, 2018.
- [21] K.-T. Kim, C. Guan, and S.-W. Lee, "A subject-transfer framework based on single-trial EMG analysis using convolutional neural networks," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 1, pp. 94–103, 2019.
- [22] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Hum. Brain. Mapp.*, vol. 38, no. 11, pp. 5391–5420, 2017.

- [23] E. Lashgari, D. Liang, and U. Maoz, "Data augmentation for deep-learning-based electroencephalography," *J. Neurosci. Methods*, vol. 346, p. 108885, 2020.
- [24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.
- [25] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Proc. Adv. Neural Inf. Process.* Syst. (NeurIPS), vol. 31, 2018.
- [26] F. Di Russo et al., "Normative event-related potentials from sensory and cognitive tasks reveal occipital and frontal activities prior and following visual events," *Neuroimage*, vol. 196, pp. 173–187, 2019.
- [27] V. Menon, J. M. Ford, K. O. Lim, G. H. Glover, and A. Pfeffer-baum, "Combined event-related fMRI and EEG evidence for temporal—parietal cortex activation during target detection," *Neuroreport*, vol. 8, no. 14, pp. 3029–3037, 1997.
- [28] X. Song, Y. Zeng, L. Tong, J. Shu, G. Bao, and B. Yan, "P3-MSDA: Multi-source domain adaptation network for dynamic visual target detection," Front. Hum. Neurosci., vol. 15, p. 685173, 2021.