

RESEARCH ARTICLE

LAD: A Hybrid Deep Learning System for Benign Paroxysmal Positional Vertigo Disorders Diagnostic

TRUNG XUAN PHAM¹, (Student Member, IEEE),
JIN WOONG CHOI², (Senior Member, IEEE),
RUSTY JOHN LLOYD MINA¹, (Student Member, IEEE),
THANH NGUYEN¹, (Graduate Student Member, IEEE),
SULTAN RIZKY MADJID¹, (Graduate Student Member, IEEE),
AND CHANG D. YOO¹, (Senior Member, IEEE)

¹School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea

²Department of Otorhinolaryngology-Head and Neck Surgery, School of Medicine, Chungnam National University, Daejeon 35015, South Korea

Corresponding author: Chang D. Yoo (cd_yoo@kaist.ac.kr)

This work was supported in part by the Institute for Information and Communications Technology Promotion (IITP) Grant through the Korean Government [Ministry of Science and ICT (MSIT)], Development of Causal AI through Video Understanding, under Grant 2021-0-01381; and in part by the National Research Foundation of Korea (NRF) Grant through the Korea government (MSIT) under Grant 2022R1A2C201270611.

ABSTRACT Herein, we introduce “Look and Diagnose” (LAD), a hybrid deep learning-based system that aims to support doctors in the medical field for diagnosing effectively the *Benign Paroxysmal Positional Vertigo* (BPPV) disorder. Given the body postures of the patient in the Dix-Hallpike and lateral head turns test, the visual information of both eyes is captured and fed into LAD for analyzing and classifying into one of six possible disorders which the patient might be suffering from. The proposed system consists of two streams: (1) an RNN-based stream that takes raw RGB images of both eyes to extract visual features and optical flow of each eye followed by ternary classification to determine left/right posterior canal (PC) or other; and (2) pupil detector stream that detects the pupil when it is classified as Non-PC and classifies the direction and strength of the beating to categorize the Non-PC types into the remaining four classes: *Geotropic* BPPV (left and right) and *Apogeotropic* BPPV (left and right). Experimental results show that with the given body postures of the patient, the system is capable of accurately classifying given BPPV disorder into the six types of disorder with an accuracy of 91% on the validation set. The proposed method can successfully classify disorders with an accuracy of 93% for the *Posterior Canal* disorder and 95% for the *Geotropic* and *Apogeotropic* disorder, paving a potential direction for research with the medical data.

INDEX TERMS Actions recognition, BPPV disorders, CNN, nystagmus, RNN (LSTM, GRU).

I. INTRODUCTION

Benign Paroxysmal Positional Vertigo is one of the most common causes of vertigo with a disorder arising from the problem in the inner ear [1]. Symptoms are repeated, brief periods of vertigo with movement, characterized by a spinning sensation upon changes in the position of the head [2]. Each episode lasts less than one minute and nausea is commonly associated. Each part of the name depicts a key feature

The associate editor coordinating the review of this manuscript and approving it for publication was Zahid Akhtar¹.

of the inner-ear disorder: *Benign* means it is not very serious, the patient's life is not in danger; *Paroxysmal* indicates that the symptom occurs suddenly and takes place within a short time; *Positional* implies that the vertigo symptom is triggered for certain postures and head movements. BPPV is considered a common public health problem [3]. The probable causes of BPPV are different lesions such as head trauma, head injury, or some surgeries related to otology, oral, and maxillofacial cases [4]. The dizziness symptom of BPPV is explained by loose calcium deposits (crystals or “ear rocks”) in what is called the semicircular canals of the inner ear [5].

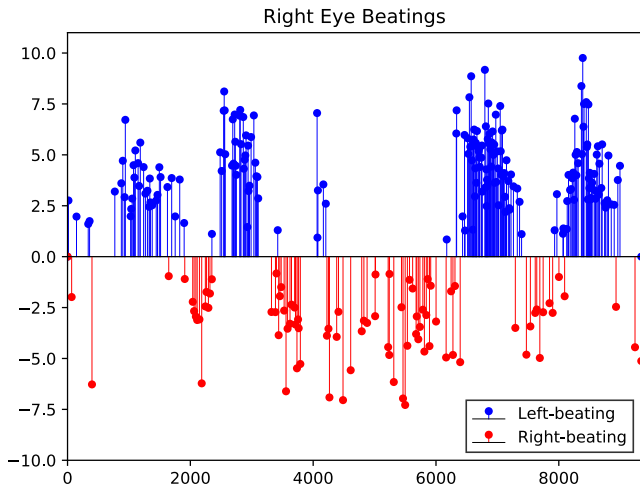


FIGURE 1. Horizontal beating detection of the right eye has been analyzed for a patient who is diagnosed to be positive to type Lt_Geo_BPPV with 9676 video frames. Red points are the right beating, blue points are left beatings. The horizontal axis is the number of frames, the vertical axis is velocity.

When the patient’s head moves, these crystals roll around the semicircular canal transmitting misleading information including sight information which causes dizziness. There are around 5 to 6 million people who suffer from this disorder annually [6], [7], [8]. In some cases, the problem of BPPV might become serious when it causes increasing chances of falling and losing balance for patients [9]. The abnormal movements of the eyes help the doctor determine the type of BPPV disorder. The symptoms of benign paroxysmal positional vertigo usually often accompany abnormal rhythmic eye movements. These behaviors include nystagmus (beating) and torsional motions of the eyes. Based on the different nystagmus movements of the eye, the doctor is able to categorize BPPV into different disorders for the appropriate treatments. Differentiating different disorders is a challenging task as it requires expert knowledge and careful observation of the patient. BPPV disorder can occur in all age groups and gender. However, it is rare for someone under 20 years to have the disorder but quite common for someone in the middle age group of 31-50 years [3].

BPPV is often benign, however, it may be dangerous in certain cases. For example, when a BPPV patient is working on a ladder or on the top roof, he may suddenly be hit by a vertiginous symptom and he may lose balance which may lead to an unwanted serious accident. If the kind of disorder is accurately diagnosed, prompt treatment can be applied so that the patient can be cured earlier and avoid the painful experience and future risks [10]. BPPV disorders are diagnosed by observing the visual motions of the human eyes in some clinical settings [11], [12], [13].

Based on any abnormal movements of the eyes, the otology experts are able to categorize each case to a specific type for further appropriate treatments. When the number of patients increases, it is a burden for medical employees. The cost

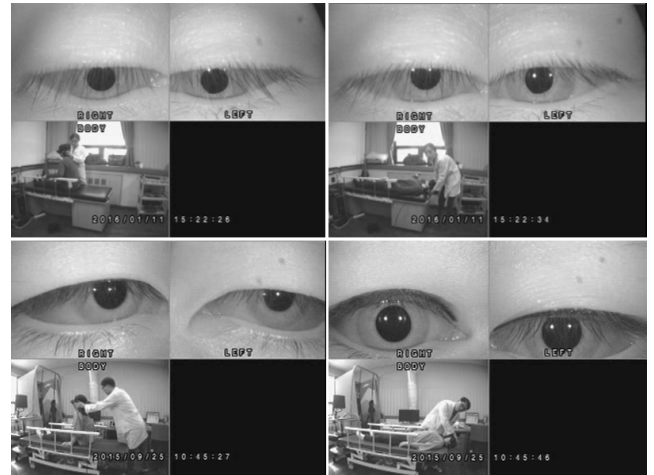


FIGURE 2. Four examples of the collected video dataset. Following the Dix-Hallpike test, [10], different poses of the patient body were conducted for the diagnosis of BPPV disorders with the visual observation corresponding to the eye’s abnormal moving. Each video contains 3 sub-videos which record the left eye, right eye, and body pose. The black area does not contain information.

of BPPV diagnosis is estimated to be approximately \$2000, and most patients suffering from this disorder may undergo unnecessary testing or other interventions which leads to the associated cost of billions of dollars per year for diagnosing BPPV alone [2]. Deep learning has made a huge impact in computer vision that includes image classification [14], [15], [16], video question and answering [17], [18], [19], object recognition [20], and natural language processing [21]. It is helping to identify, classify, and quantify patterns in medical images. In [22], Inception V3 is used to extract new knowledge from retinal fundus images and infer multiple cardiovascular risk factors with the high area under the receiver operating characteristic curve (AUC). In [23], the fine-grained variability in the appearance of skin lesions has been effectively captured by using CNN with manipulated transfer learning to detect skin cancer and achieve performance on par with the clinical experts. Recent studies demonstrate the potential applications of deep learning for understanding human actions in the videos [24], [25], [26], [27]. These researches show the promising capability of deep learning in capturing the torsional nystagmus patterns in the fine-grained motion of the eyes for BPPV analysis. In this paper, a deep learning system referred to as “Look and Diagnose” (LAD) is capable of classifying BPPV disorder into six types of the most common BPPVs including a geotropic lateral canal (left and right), apogeotropic lateral canal (left and right), and posterior canal types (left and right) based on the video of both eyes and seven postures of the patient. LAD is capable of classifying different types of beating (nystagmus) including torsional and horizontal beating to effectively classify BPPV disorders. To develop LAD, we have collected a large dataset of both eye movement videos of BPPV disorders. More details are shown in Section IV.

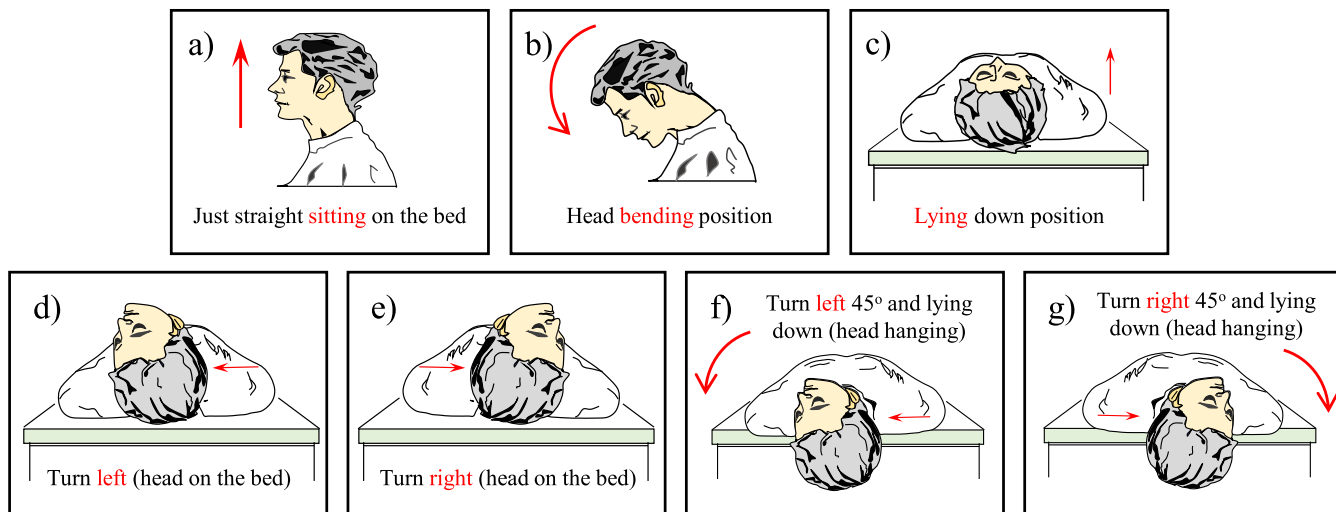


FIGURE 3. A normal procedure of BPPV diagnosis which collects video data of the eyes associated with the seven positions of the patient’s body: (a) the patient is sitting straight, (b) head bending while sitting position, (c) laying down while patient keeps looking straightly into the ceiling, (d) when patient laid on the bed, perform turning head left quickly, (e) when patient laid on the bed try turning head right quickly, (f) from the sitting position, try turning head left 45° lying down and make sure the head is hanging out of the bed, and (g) from the sitting position, turning head right 45° lying down and make sure the head is hanging out of the bed. Performing actions *b, c, d, e* often indicates the “lateral head turns” test while the action *a, f, g* are intended the “Dix-Hallpike” test.

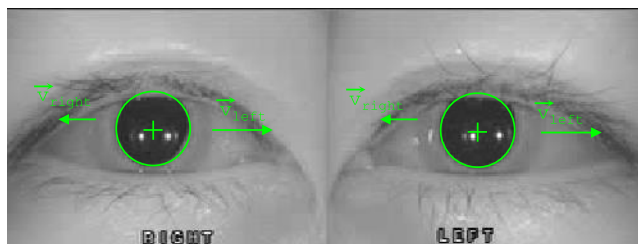


FIGURE 4. A sample of our video data. BPPV disorder is specified by the natural nystagmus of the people’s eyes along with the featured motions such as torsional and spontaneous beating movements. Horizontal beatings are measured and detected by the proposed algorithm to diagnose the lateral canal BPPV types, while torsional motions are captured by the RNN (GRU) network.

II. RELATED WORKS

A. BPPV ANALYSIS

Nystagmus examination is a crucial step to diagnose BPPV disorder [1]. Peng et al. investigated different maneuvers to provoke the nystagmus for diagnosis and the BPPV can be treated effectively by applying repositioning moving [10]. Previous studies show that horizontal (lateral) canal type (5% to 15% of BPPV cases) is less common than the Posterior canal type (85% to 95% of BPPV cases) [2].

Slama et al. [28], [29] used a multilayer neural network (MNN) and the recorded parameters from Video Nysta Geographic (VNG) data to analyze the nystagmus to diagnose whether one person has a vestibular disorder or normal. Nystagmus signal from Videonystagmography (VNG) device has been analyzed by CNN-based method to classify two classes of vestibular disorder [30]. Lim et al. [31] introduced a more complete approach that does not just only classify the nystagmus, but also diagnoses the final BPPV class using various ad-hoc techniques.

Lim et al. recorded the tracking videos of eyes at ten different postures of the patient. These videos have been processed to get the cumulative transient velocity of the eye movement by simply subtracting eye coordinates for every video frame. The extracted scalars have been formed into a grid image for each video and fed into a standard CNN classifier. Lim et al. is based on hand-crafted features for diagnosing BPPV disorder, which has limited their performance.

In contrast to [31], we collected videos from our clinical warehouse with seven postures of patient (Fig. 3, Fig. 2). A bidirectional GRU-based network was used to successfully learn end-to-end the torsional movements and nystagmus simultaneously of the two eyes from the given video input.

A consecutive feature-based classifier captures the horizontal beating of the eyes and recognizes the remaining four horizontal (lateral) canal BPPV disorder classes including geotropic BPPV (left and right side) and apogeotropic BPPV (left and right).

B. ACTION RECOGNITION

Video understanding is an attractive and challenging task. Some earlier works tried to extract the handcrafted features of the video via improved Dense Trajectory (iDT) and combine it with low-level video descriptors for video representation [24], [32], [33]. Deep learning has made a significant impact on human action recognition [34], [35], [36], [37], [38], [39], [40]. Two-stream networks are capable of extracting relevant features regarding the appearance and motion information from RGB and optical flow field for accurate action recognition [25], [27], [41].

Another approach for action recognition uses the 3D-CNN networks to directly capture motion clues in the video by

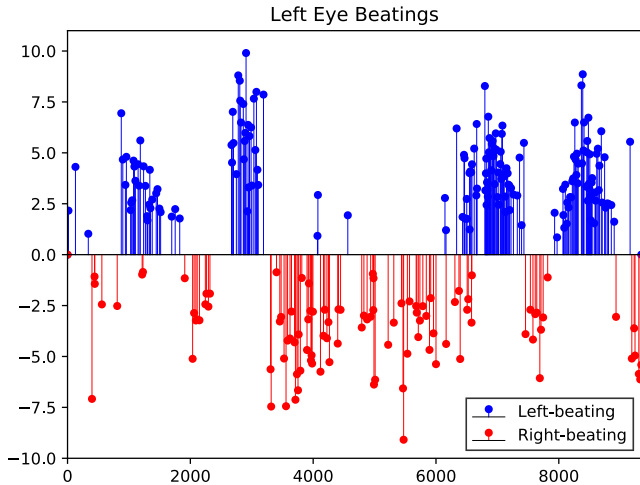


FIGURE 5. Horizontal beating detection of the left eye has been analyzed for a patient who is diagnosed to be positive to type Lt. Geo. BPPV with 9676 video frames. Red points are the right beating, blue points are left beatings. The horizontal axis is the number of frames, the vertical axis is velocity.

using 3D convolution without using optical flow field [42], [43]. In this work, a two-stream system is adopted to extract the motion features such as the torsional nystagmus of the eyes for accurate disorders classification.

C. EYE PUPIL TRACKING

Traditional computer vision methods use edge features, intensity gradient distribution and intensity thresholds to detect the eye pupil [44], [45]. A more advanced feature-based technique uses edge segment selection and conditional segment combination for pupil detection with high accuracy [46]. Recently, a deep learning-based network is incorporated to robustly detect pupils even under extreme conditions with reflections and occlusion by eyelids [47], [48].

In this work, the pre-trained CNN model for localizing the pupil proposed by [48] is adapted, and the pupil position is monitored in the proposed system for classifying horizontal beating so that four types of lateral canal BPPV disorders are categorized. To this end, an algorithm is proposed to process the pupil location to detect when the beating happened and measure the speed of horizontal beating for lateral canal BPPV.

III. METHODOLOGY

This section describes the details of the proposed BPPV disorder diagnosis system referred to as LAD which takes as input the video recording of eye movement of both eyes with posture labels. Seven postures labels aligned with the eye movement are provided. Fig. 3 shows the seven postures that include sitting, head on the bed turned left, head on the bed turned right, head hanging 45° to the left, head hanging 45° to the right, lying down position, and head bending a sitting position.

Based on these positions, the system learns to capture the relevant features of both eyes from RGB video and optical

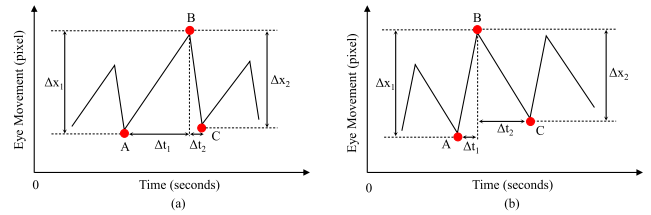


FIGURE 6. The horizontal beating is determined by slow phase velocity (SPV) and fast phase velocity (FPV). Fig. (a) illustrates the eye has left-beating. In this case, the eye move from point B to point C faster (decreasing the pixel) than the eye moves from A to B. Whilst Fig. (b) shows the right-beating behavior of the eyes where they move from point A to point B faster than when it moves from point B to point C. The behavior of these beatings is contradicted. Here, Δx_1 indicates the distance when the eye travels during the AB in the direction of pixel increasing and Δx_2 is the distance when the eyes move toward the decreasing pixel direction. Similarly, we can determine the beating up and beating down where the eyes have the beating direction vertical. The beating is counted if the beating velocity is above a threshold.

flow output of the optical flow CNN and based on a bi-directional gated recurrent unit (BiGRU) framework categorizes the input video into two types of posterior canal BPPV (PC) and other (not PC). If the video is classified as not PC type, then the beating feature of the eye is extracted to categorize it into one of four remaining BPPV disorder types.

Algorithm 1 Beating Detection (pseudo code), Fig. 6

- 1: **Input:** Detected x coordinates of pupil.
- 2: **Output:** Beating velocity, frame ID beating occurred.
- 3: Initialization: $counter_1 \leftarrow 1, counter_2 \leftarrow 1, beat_{left} \leftarrow NULL, beat_{right} \leftarrow NULL$
- 4: **for** frame $i=2:N$ **do**
- 5: $flag_1 \leftarrow \text{sign}(x[i-1]-x[i-2])$
- 6: $flag_2 \leftarrow \text{sign}(x[i]-x[i-1])$
- 7: **if** $flag_1 > 0$ and $flag_2 > 0$ **then**
- 8: Increase $counter_1$
- 9: **else if** $flag_1 < 0$ and $flag_2 < 0$ **then**
- 10: Increase $counter_2$
- 11: **else if** $flag_1 > 0$ and $flag_2 < 0$ **then**
- 12: Calculate slow phase velocity: $v_{left} \leftarrow \frac{\Delta x}{counter_1}$
- 13: **else**
- 14: Calculate fast phase velocity: $v_{right} \leftarrow \frac{\Delta x}{counter_2}$
- 15: **end if**
- 16: **if** $v_{left} < v_{right}$ **then** store $beat_{right}$ **else** store $beat_{left}$, reset counters to 1.
- 17: **end for**

A. CLASSIFYING POSTERIOR AND LATERAL TYPES

1) POSTERIOR CANAL DISORDER

It is specified by the torsional movement of the eye triggered under certain postures of the patient. The posterior BPPVs are categorized into two types: the first one is the “left posterior canal” type (Lt_PC_BPPV) specified by nystagmus of torsional clockwise along with a slight up-beating when lying with head hanging, and the second one is “right posterior

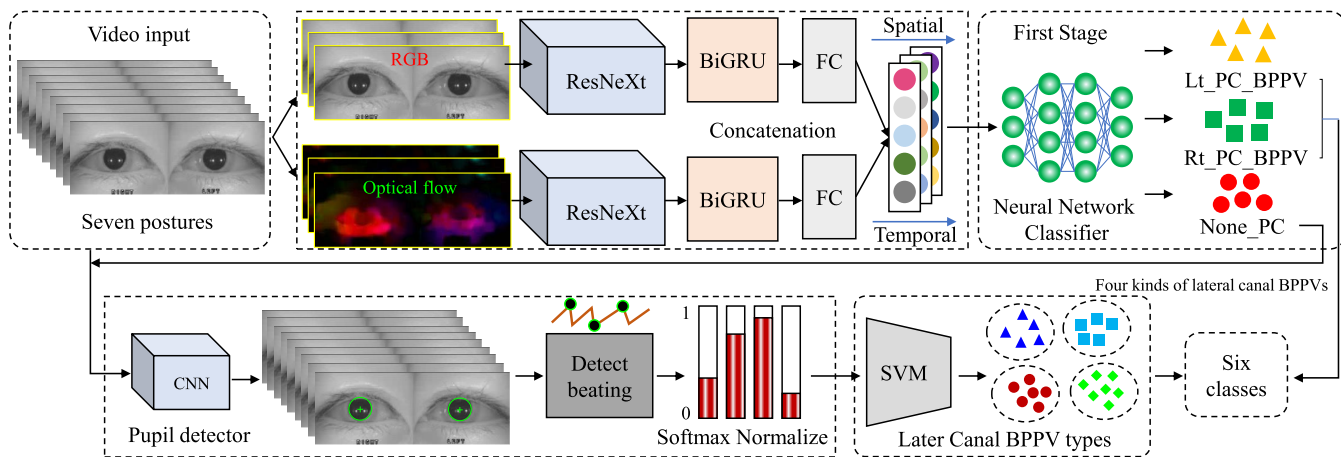


FIGURE 7. The whole architecture for detecting six BPPV disorders. In the first stage, the input video is categorized into three raw classes “posterior types” and “not posterior types”. After that, a pupil detector based on CNN is applied to track the eye pupils and detect the beating that happens in the second stage.

canal” type (Rt_PC_BPPV) with the nystagmus of torsional counter-clockwise alongside with a slight up-beating when lying with head hanging (shown in Fig. 8). For a given video input, we first classify whether it belongs to posterior canal types or not. To classify posterior BPPVs, we use the recurrent neural network with features extracted from a pre-trained CNN for two eyes of the patient at five positions as in Fig. 3: c) d) e) f) g). At each position, the eye’s motion is captured by a bidirectional GRU followed by a fully connected layer with 128 neurons. The outputs of five positions are concatenated to a 640-dimension vector and fed into a fully connected layer and a final softmax layer for classification (Fig. 7).

2) LATERAL CANAL DISORDER

BPPV with the horizontal beating of the eyes (shown in Fig. 4) should be distinguished from posterior canal BPPV (PC) which is associated with torsional nystagmus of the eye. Based on the beating direction of the patient eyes corresponding to his/her body posture, it is categorized into geotropic and apogeotropic BPPV (next section). According to the Barany Society [49], the horizontal canal (HC) BPPV is categorized into two different types Canalolithiasis (HC-geo) and Cupulolithiasis (HC-apo) types for each ear (left and right side). So, the total types of HC-BPPV are 4 types: 1) Right ear HC-geotropic (canalolithiasis), 2) Right ear HC-apogeotropic (cupulolithiasis), 3) Left ear HC-geotropic (canalolithiasis), and 4) Left ear HC-apogeotropic (cupulolithiasis) [49]. To classify the different types of geo and apogeo BPPV, LAD extracts the number of beatings and the velocity magnitude of these beatings. And this information is utilized to form a final 20-dimension features vector for each given video input.

B. CLASSIFYING GEOTROPIC AND APOGEOTROPIC TYPES

Geotropic BPPV includes: left- and right- lateral canal BPPV, we denote them as “Lt_Geo_BPPV” and

“Rt_Geo_BPPV”, respectively. In the same way, we have the **apogeotropic** BPPV: “Lt_Apogeo_BPPV” and “Rt_Apogeo_BPPV”. The definition of these disorders is summarized below. We performed experiments using LSTM or GRU for these types of BPPVs and found that it is not better than random guess indicating that RNN is hard to learn the horizontal beating. All four different types of the lateral canal are characterized by horizontal nystagmus, and end-to-end learning of CNN+RNN did not produce satisfactory classification results. Horizontal nystagmus occurs within a certain moment in time (20s-30s), and “strong beating” in each lesion side is the most important indicator in classifying these BPPV disorders (specialists perform diagnosis by looking at particular characteristics of the horizontal nystagmus). In this case, the CNN+RNN model failed to capture the “strong beating” signal. The proposed **Alg. 1** detects strong beating by way of measuring the speed of the beating, which is the most important indicator for classifying different lateral canals.

In the second stream, LAD is learned to extract the different types of features to categorize horizontal beating and classify the beatings into one of four types of horizontal BPPV disorders.

1) GEOTROPIC BPPVs

There are two geotropic BPPV types that are defined as follows: Left geotropic canal BPPV composes the geotropic nystagmus and stronger intensity on the left side; while right geotropic canal BPPV includes the geotropic nystagmus and accompanies the stronger intensity on the right side. Both of the cases are considered from the patient’s perspective for the postures.

2) APOGEOTROPIC BPPVs

Two types of Apogeotropic BPPV are differentiated conversely with the geotropic types: Left apogeotropic canal BPPV with apogeotropic nystagmus and stronger intensity

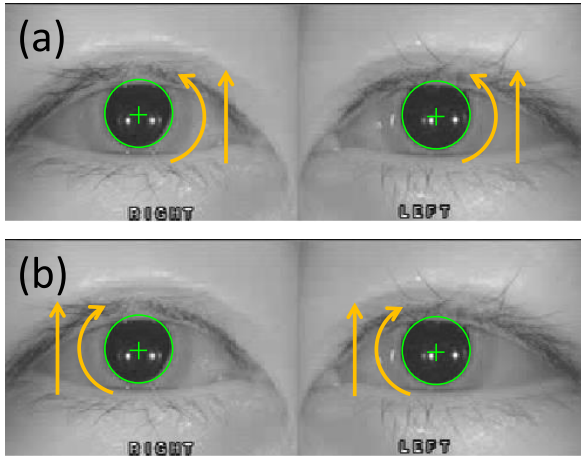


FIGURE 8. The posterior canal BPPV types with (a) Rt_PC_BPPV specified by a counterclockwise torsion and sometimes accompany with slight up beating, and (b) Lt_PC_BPPV type with clockwise torsion and up beat. These featured movements of the eyes have been learned successfully by the proposed system which is much more robust than the conventional template-matching -based method as in [31].

on the right side; right apogeotropic canal BPPV with apogeotropic nystagmus and stronger intensity on the left side.

C. EYE BEATING DETECTION

Normally, a video is recorded with the full seven actions of the subject patient with the support of a skillful doctor. In some cases, “head bending” may not be included. We detect horizontal beating for acquiring the visual features of the eyes to classify four types of lateral canal BPPV disorder. We can use different available methods to detect and track the eye pupil from the classical techniques that use feature-based methods such as Hough Transform [45] or the more recent advanced technique PuRe [46]. Recently, deep learning is an interesting approach for tracking the pupil [47], [48]. Here we detect eye pupils with a CNN-based network that was inspired by Shaharam et al. [48]. In our dataset, as described in section IV, to trigger and observe the clear symptom of BPPV disorders, the clinical experimental settings are done in the low light condition to record the videos of the two eyes. In addition, we find that the size and shape of the pupils varied for different patients, and there are cases the eyelid covers the pupil partly or half for which the traditional pupil detector failed to detect correctly the pupils (contains lots of false positives). In practice, we find that the neural network-based CNN detector [48] performs more robustly in detecting the pupil compared to the traditional method such as Hough Transform, furthermore, CNN is supported to run on GPU with high speed in processing (~120 FPS for a single GPU Titan Xp NVIDIA, compared to the traditional detector Hough with ~70 FPS).

The outputs are the coordinates of the eye (x, y) and the radius of detected pupils r . This information of coordinates is utilized to calculate the velocity of the eye movement for

beat detection. With each given video, the output should be the beating with the magnitude of beating specified by the velocity magnitude stick to the frame ID that the beating is detected. With each of the positions of the patient, we have the total number of beats left, and right, and the corresponding velocity magnitude. This information is stacked to form the final feature of each video to distinguish four types of BPPVs (Geo and Apogeo types). Algorithm 1 detects the horizontal beating using the coordinates of the eye pupil acquired from every video frame.

SPV and FPV are computed by Eq. 1. The coordinates of the eye’s pupil (x, y) include x for the horizontal axis and y for the vertical axis. For the horizontal canal BPPV, we focus on the horizontal component x only. Algorithm 1 illustrates how to detect one beating that occurred with the coordinate x of the eye. The beating velocity is computed by the following equation (Fig. 6):

$$v = \frac{\Delta x}{\Delta t}, \tag{1}$$

where Δx is the distance the eye travels within the space of the video frame during the time step Δt . The beatings for two eyes of the patient are scattered in Fig. 1 and Fig. 5, the behavior of both eyes is almost the same. In some cases, if one of the eyes closes then we can still get the information from the other one. Fig 1 shows the horizontal beating over time for a video of a patient who is positive with left lateral canal BPPV.

In the region from frame 1900 to frame 2200 when the patient turned his head right, the right beating is dominated (red points). A few moments later, when he turned left (frame 2500 to frame 3000), the left beating appeared (blue points).

D. FEATURES FROM DETECTED BEATING

For each left and right pair of beating, a softmax (Eq. 2) is applied to normalize the raw beating number to a range between 0 and 1. The pair of average beating velocity magnitudes are also normalized in the same way. For horizontal canal disorder, we care for the beating on the positions in Fig. 3 *d, e, f, g*. With four positions of the patient, we have a total of **eight** normalized values of the left and right beating numbers. The binary value 0 for the left side and 1 for the right side, denotes the side that has the maximum total beating, and performed a softmax normalization for the pair of the max total beating on the left side and the max total beating on the right side as the following equation:

$$\text{Softmax}(\mathbf{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^2 \exp(x_j)}, \tag{2}$$

where $i \in 1, 2$ and $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$. After this step, we have three more features for positions *d, e*. Now apply the same way for the average beating velocity for both sides, we have three more features for positions *d, e* (lateral head turn test). Similarly, *six features are extracted for the position f, g* when the patient lying on the bed (with the head hanging left and right with the Dix-Hallpike maneuver). For the position

TABLE 1. Classifying the torsional nystagmus (posterior canal types).

Reference	Feature extraction	Classifier	Be able to learn End-to-End
Lim <i>et al.</i> [31]	Hand-crafted features (Template Matching)	CNN	✗
The proposed LAD	CNN (pre-trained on ImageNet dataset)	CNN+BiGRU	✓

TABLE 2. Classifying the horizontal nystagmus (lateral canal types).

Reference	Feature extraction	Eye tracking method	Classifier	Detect beating
Lim <i>et al.</i> [31]	Sum of the transient velocity (all frames)	Hough Transform	CNN	✗
The proposed LAD	Velocity and count of detected beatings (filtered frames with no beating)	CNN	SVM	✓

TABLE 3. Comparison of the recent works in analysis of nystagmus. The performance is reported based on the accompanying dataset. The abbreviations are as follows: PCA: Principle Component Analysis, MNN: Multilayer Neural Network, VNG: Video-nystagmography, FLD: Fisher Linear Discriminant, CNN: Convolutional Neural Network, VOG: Video-oculography, F-CNN: Fully Convolutional Neural Network, BiGRU: Bi-directional Gate Recurrent Unit.

Reference	Published year	Feature dataset	Classifier	Classes	Body pose video	Overall accuracy.
Ben <i>et al.</i> [50]	2017	Temporal, frequency	PCA-MNN	3	✗	-
Ben <i>et al.</i> [29]	2018	Clinical VNG features	FLD-MNN	2	✗	-
Lim <i>et al.</i> [31]	2019	Eye tracking video	Template matching + CNN	8	✗	79.8 %
Ben <i>et al.</i> [30]	2019	Nystagmus signal	CNN	3	✗	-
Yiu <i>et al.</i> [51]	2019	VOG sequence	F-CNN	2	✗	-
Ben <i>et al.</i> [28]	2020	VNG sequence	FLD-MNN	2	✗	-
The proposed LAD	2022	Eye tracking, body video	Hyrid CNN+BiGRU	6	✓	91.0%

of head bending (Fig. 3b), and lying down (Fig. 3c), the beating direction is important, the magnitude of beating is not considered when diagnosing lateral canal BPPV types.

Finally, we form a 20-dimension vector extracted from a video that contains information on the horizontal beating of both affected sides of the patient. These feature vectors are used as input to a classifier such as a deep neural network or an SVM for classification. The experimental results and analysis are described in the next section. We also conducted the experiments with the method in [31] that is most relevant to our work on the nystagmus analysis for BPPV.

IV. EXPERIMENTS

A. DATA COLLECTION

The videos were recorded in *.avi format with the resolution of 240 × 320 and at 30 FPS, each video includes three sub-videos: left eye, right eye, and whole patient’s body scene. As shown in [52], to observe the symptom of the eyes clearly, the diagnostic tests were conducted in a dark environment. There are two popular positional tests that are used generally to diagnose BPPV disorder which are described in more detail below. The room is set up with a low light condition to trigger BPPV symptoms in patients. The patient is staying on the bed. There is one doctor standing by side to perform Dix-hallpike maneuvers and lateral canal tests. There is one camera located 2.3m far from the patient to capture the whole scene of both patient and doctor.

The first test is the Dix-Hallpike maneuver which is used to diagnose the type of posterior canal BPPV [10]. This test is performed independently for both the left and right sides. For the left side, the patient (equipped with a goggles camera attached to his head) is seated straight on the bed and a doctor will help him to turn his head 45° to the left and quickly lie

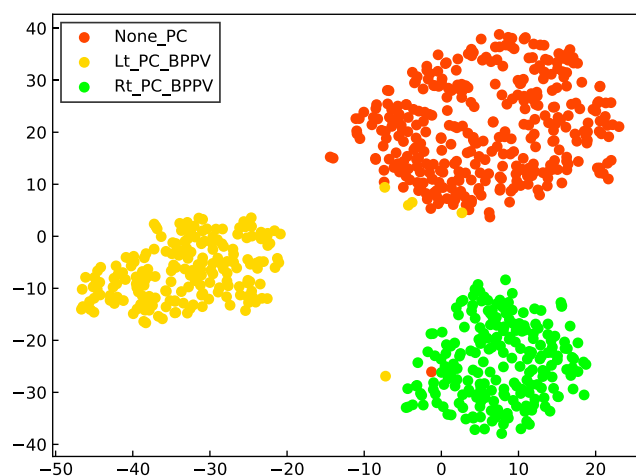


FIGURE 9. t-SNE plot for the first round, classifying video into three kinds of BPPV groups: Lt_PC_BPPV, Rt_PC_BPPV, and None_PC. The plot is scattered into the 2D plane for 746 data points (the whole dataset). It is best viewed in color.

down so that the patient’s head is hanging left out of the bed (Fig. 3f). The right side test is similar to the left side but now the head of the patient is turned 45° to the right before lying down in a hanging position (Fig. 3g).

Keeping each side test duration for less than thirty seconds and the eyes of the patient are observed for any nystagmus that occurred during this period. In the Dix-Hallpike test, the vertical-torsional nystagmus is often observed either clockwise or counter-clockwise direction which specified the left posterior canal or right posterior canal BPPV type, respectively.

The second test performed on the data collection is the lateral canal test [2]. This test is done by keeping the patient

lying (head is on the bed) with the eyes looking straight to the ceiling and conducting two maneuvers. For the left side, quickly turn the patient's head toward the left (Fig. 3d) For the right side, turn the patient's head quickly to the right (Fig. 3e). After each of two of these maneuvers, the eyes of the patient are observed within less than thirty seconds to see if any horizontal nystagmus has occurred during this period.

In the lateral canal test, four types of BPPV are clearly observed including left and right geotropic lateral canal BPPV and left and right apogeotropic lateral canal BPPV. The patient's body (captured by a remote camera) and two eyes (captured by goggles) were performed simultaneously and stacked to a video with three sub-videos for the left eye, right eye, and the body (to see the scene later). Each video belongs to a unique patient, ranging from 10 to 70 yrs old, including both genders.

The two physical tests are consecutively recorded resulting in a unified video with a duration of approximately three minutes and a size of 15-20 MB on average. By observing the behaviors of the patient eyes in all positions we tested above, two otology experts in a hospital analyze and classify which kind of disorder the patient belongs to. The recorded videos with subtle symptoms, low quality, or incomplete have been removed from the cohort. A total of 746 data from patients with six classes of BPPV disorder with 406 data for posterior canal types and 340 data for lateral canal types were used in the experiments.

Table 4 shows that posterior canal BPPV is more common and accounts for a proportion of 26% and 29% for the left and right types, respectively. Table 3 illustrates a comparison of different works on various data collections in BPPV analysis. Our dataset is different from previous datasets: it includes movements of both eyes that are aligned with the different postures of the patient (following the Dix-Hallpike test procedure).

B. TRAINING DETAILS

A pre-trained CNN ResNeXt32 \times 48d [53] is used as a feature extractor for both spatial and temporal streams which results in a 2048-dimension vector for each frame of the two eyes. We tried with another feature extractor Inception V3, however, the recognition accuracy is behind 5% compared to the ResNeXt32 \times 48d. In the part of classification for posterior BPPVs, we utilized a bidirectional GRU to learn the eye motions in five corresponding postures of the patient.

In this part, each GRU network outputs a vector of 128 dimensions and is stacked to form a final 640-dimensions vector. A fully connected layer (512 neurons) and a softmax layer (3 neurons) were added to classify the given video into one of 3 classes: None_PC, "Lt_PC_BPPV" and "Rt_PC_BPPV". All of these networks used ADAM optimizer [54] with default hyper-parameters for training and cross-entropy loss for classification. The learning rate was set to 0.001.

In the classification part for four types of BPPVs, each video was extracted to a 20-dimension vector and then fed

TABLE 4. The collected data was split into train and test sets with a ratio of 90% and 10%, respectively. "Per class rate" refers to the proportion of data for each class in the whole data.

Disorder types	# Train sample	# Valid sample	Per class rate
Lt_Geo_BPPV	73	9	11 %
Rt_Geo_BPPV	78	9	12 %
Lt_Apo_BPPV	84	10	13 %
Rt_Apo_BPPV	69	8	10 %
Lt_PC_BPPV	173	20	26 %
Rt_PC_BPPV	191	22	29 %
Total	668	78	100 %

into a linear-kernel SVM classifier for classification [55]. The use of SVM was a designer's choice as SVM performed similarly to FC+softmax and SVM was a simpler choice requiring less memory.

C. EVALUATION METRICS

To evaluate performance, the following metrics were used: (1) accuracy, (2) precision, (3) recall, and (4) F1-score. Accuracy is computed by the proportion of correct predicted class over total data. Precision shows the proportion of predicted data that are correctly classified. Recall gives the proportion of actual samples that are correctly classified.

The F1-score determines the performance of the model to classify all data, which makes a balance between precision and recall and gets its best value at 1 and worst score at 0. These standard measures are computed and presented by the following standard equations:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (6)$$

where TP, TN, FP, and FN denote the True Positive, True Negative, False Positive, and False Negative, respectively. In the total 746 data samples, train and valid set are split randomly with a ratio of 90%/10% three times.

The aforementioned metrics are calculated for three validation sets and averaged. The split procedure follows that used in obtaining the splits for JHMDB/ HMDB dataset for action recognition [56], [57]. Instead of a 7:3 split of HMDB, we used a 9:1 split for the evaluation.

D. RESULTS

We show that the four lateral canal types of Geotropic and Apogeotropic BPPV disorder can be well classified based on the proposed method. Fig. 10 visualizes the feature vectors of lateral BPPV types in 2D plane with t-SNE visualization [58]. It clearly shows that these vectors are clearly separable and belong to four separate groups.

These feature vectors are constructed by the measurement of horizontal beating velocity alongside magnitude and the

TABLE 5. Input data have been roughly categorized into three groups using BiGRU: Not posterior canal (None_PC_BPPV), left posterior canal (Lt_PC_BPPV), and right posterior canal BPPV (Rt_PC_BPPV). It is performed on three different splits validation sets and takes the average. Then, the None_PC_BPPV is classified into the rest of the four classes with lateral canal BPPV types.

BPPV types after the 1 st stage	Accuracy	Number of samples
Lt_PC_BPPV	0.90	20
Rt_PC_BPPV	0.94	22
None_PC_BPPV	0.94	36

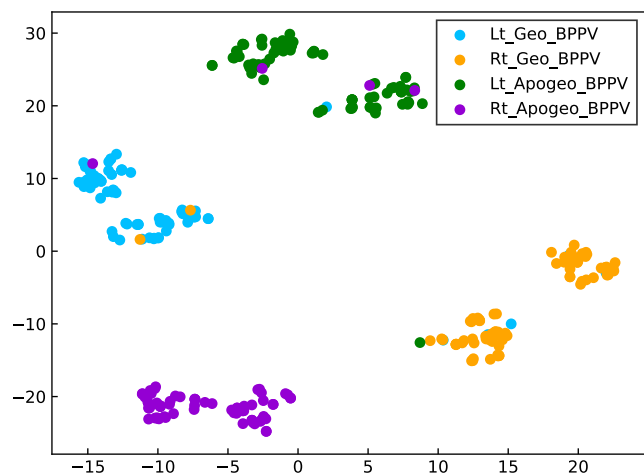


FIGURE 10. t-SNE plot for 4 kinds of BPPV disorder: geotropic (left, right) and apogeotropic (left, right) classes (plotting for a total of 340 data points for lateral canal types). It is best viewed in color.

TABLE 6. Accuracy of two stages and the final process. Step 1, data are categorized into posterior canal types (Lt_PC_BPPV, Rt_PV_BPPV) and None_PC_BPPV. Step 2, the None_PC_BPPV data are classified into four types of lateral canal BPPVs. And finally, the accuracy for the final predicted labels is calculated.

Type	Posterior type	Lateral type	Overall accuracy
Accuracy	0.93	0.95	0.91
Number of videos	42	36	78

beating intensity which are the crucial clues for distinguishing the horizontal canal BPPVs. Table 6 shows the validation accuracy of 95% for lateral canal type, 93% for posterior canal types, and 91% for overall. For the posterior canal BPPV, the classification results are shown in Fig. 9, Table 6, and Table 5 with three classes: Lt_PC_BPPV, Rt_PC_BPPV, and None_PC_BPPV, showing that the BPPV types with rotational features are well extracted.

To visualize how the “gated recurrent unit” (GRU) differentiates different posterior canal BPPV types, the features from the top layer of the model in Fig. 7 are extracted. The classification results and t-SNE plot demonstrate that the aforementioned featured motion of the eyes has been successfully captured by the bidirectional GRU model (BiGRU) to distinguish posterior canal types (torsional movements) and lateral canal types (horizontal beatings). Overall, the whole system classifies the videos in the validation set with 0.91,

TABLE 7. Comparison of the baselines and the proposed method in terms of validation accuracy in our dataset, † indicates that we implemented the method [31] in our dataset. Non-Information Rate (NIR) indicates the most proportional class rate in the collected dataset.

Method	Overall Accuracy (%)
NIR (baseline)	29.0
Lim et al. [31]†	48.7
Proposed LAD	91.0

TABLE 8. The classification results for six classes of BPPV disorders. The results are performed on three different splits validation sets and are taken average. LC stands for “lateral canal” while PC stands for “posterior canal”.

Type	Disorder type	Precision	Recall	F1 score	# Test samples
LC	Lt_Geo_BPPV	0.92	0.85	0.88	9
	Rt_Geo_BPPV	0.81	0.89	0.84	9
	Lt_Apo_BPPV	0.94	0.97	0.95	10
	Rt_Apo_BPPV	0.96	0.88	0.91	8
PC	Lt_PC_BPPV	0.92	0.90	0.91	20
	Rt_PC_BPPV	0.93	0.94	0.93	22
Average / Total		0.91	0.90	0.90	78

0.90, and 0.90 for precision, recall, and F1 scores as shown in Table 8, respectively.

Table 7 compares the baselines and our method. It shows that method in [31] has an accuracy of 48.7% which is much lower than that of our 91.0% on our dataset. Low resolution, low-lighting conditions, and partially obscured iris make it a challenging problem, and the simple template-matching based-method is not robust enough to capture the torsion in the iris. It also shows that our algorithm is far superior in comparison to the temple-matching method [31] which can never be robust to sporadic eyelid closure.

Alg. 1 detects frames with beating, which is important information to specialists for diagnosing lateral canal BPPV, it works as the attention mechanism filters out all frames with no beating, and the method in [31] doesn’t have this capability. The pupil detection is performed independently to mimic how experts are performing the beating (nystagmus) detection.

Table 3 shows the recent works in BPPV research. Each reference has a separate dataset and a different number of classes for nystagmus analysis. The most related data to our work is the work of Lim et al. [31]. They collected data from ten different positions of the patient, however, they failed to apply RNN to classify the disorders and used the handcrafted feature as the feature descriptor. In contrast, our dataset contains 7 postures of the patient, applied successfully an RNN (with a bidirectional GRU) for classification of BPPV disorders, and proposed an algorithm that detects the horizontal beating that can filter out all irrelevant frames that have no beating, which results in the more accurate information for classifying the torsional movements as well as the horizontal nystagmus in the given video input.

E. LEARNED FEATURE VISUALIZATION

Fig. 11, 12, and 13 visualize the features of ten different samples from three classes that have been learned by using

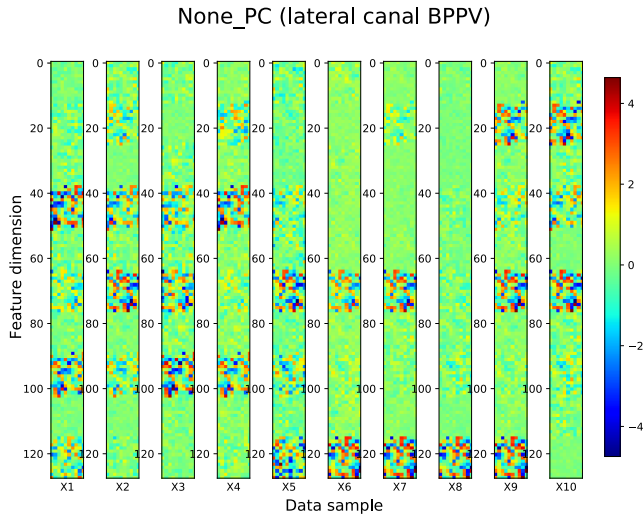


FIGURE 11. Visualizing the learned features of the None_PC class.

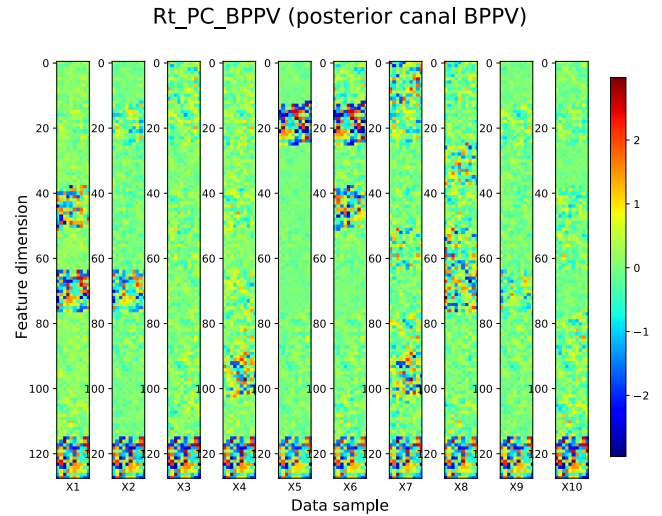


FIGURE 13. Visualizing the learned features of the Rt_PC_BPPV class.

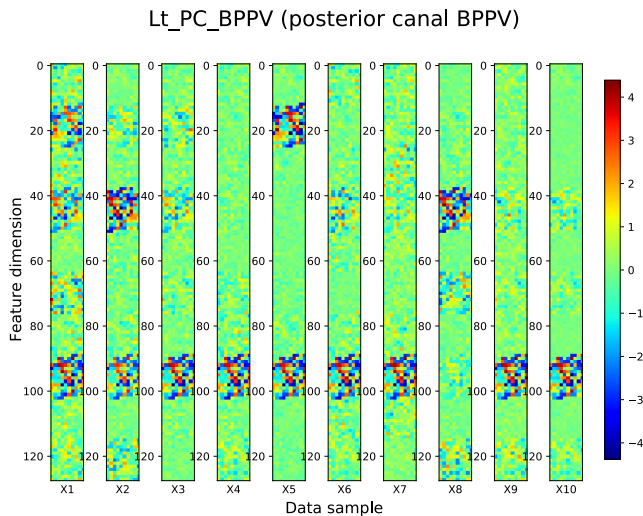


FIGURE 12. Visualizing the learned features of the Lt_PC_BPPV class.

the top layer that is trained in the proposed model. The top layer of the proposed model produces 1280-dimensions feature vectors that have been reshaped to 2D feature maps (10 × 128) for visualization. It shows that class “None_PC”, “Lt_PC_BPPV” and “Rt_PC_BPPV” have been extracted with the discriminative representations, indicating that the proposed model with RNN network has successfully learned the behaviors of torsional (Lt_PC_BPPV and Rt_PC_BPPV) and horizontal (None_PC) movements of the eyes. These visualizations also strengthen the t-SNE plot in Fig 9 that three BPPV types have been well clustered.

It is important to categorize well the video input in the first stage so that the next stage can use its results for the next stage to diagnose the rest of the BPPV types. For comparison with the full features-based method in [31] that is closest to our work, we reimplemented that technique with

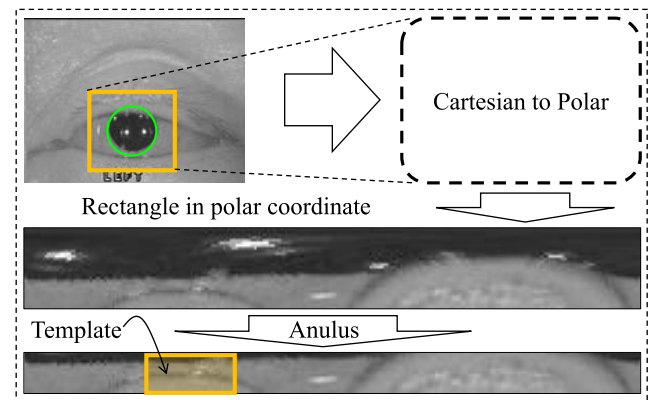


FIGURE 14. The process to convert an annulus in Cartesian coordinate to a rectangle in a polar coordinate that is described as in method [31].

template-matching for torsional categorization in our dataset (Fig 14). The comparison results are reported in Table 7.

V. CONCLUSION

This paper proposes a deep learning-based system to support the doctor in automatically diagnosing BPPV disorder effectively based on the visual motions of the patient eyes in the standard maneuver tests. Abnormal eye behaviors are triggered and these features are captured and classified into six different types of BPPV. The torsional movements of the eyes are successfully captured by a BiGRU network, while the horizontal nystagmus is distinguished by the normalized beating feature that is extracted from the eye pupil’s horizontal coordinate.

By detecting accurately the horizontal beating that occurred in a specific timestamp of the whole video, the final features of the input video are extracted to disorder types. A two-stream deep architecture network is constructed to extract different types of features for accurate classification. The proposed method can successfully classify disorders with

an accuracy of 93% for the *Posterior Canal* and 95% for the *Geo* and *Apogeo* BPPV, respectively.

VI. DISCUSSION AND FUTURE WORKS

Our work focuses on the application of artificial intelligence to investigate the BPPV disorders in medical data which can be observed by visual movements in the recorded videos. This can aid the doctors significantly in the context that the number of patients increases and hence the video data collected arose every year. The recent advances in deep learning such as CNN demonstrate the very powerful capabilities in learning the image and video representation in the carefully collected dataset [59] which is also proved to be beneficial in the custom dataset as in our work.

Our work has some limitations such as the feature of the patient body when integrates into the eyes movement features do not give the desired performance. Therefore, currently, we use labels that mark the exact posture of the patient for each moment in the video. To make the whole process a fully automatic diagnosis, the feature extractor on the human body should extract more accurate subtle details like when the head turns left or right which is crucial in diagnosing BPPV disorders.

The more recent framework Vision Transformer (ViT) [60] has emerged as a promising approach that outperforms CNN in some tasks in visual recognition. It is a potential backbone for improving the BPPV disorders automatic process. ViT is a powerful model but heavy computation, especially in the video dataset. We let it for future work.

ACKNOWLEDGMENT

The authors would like to thank Jaewon Lee, Minseong Yoon, and Jeeun You for their helpful advice and strong support during the data collection and reviewing of the article. They would like to thank KAIST U-AIM lab members for labelling the data and data checking. (*Trung Xuan Pham and Jin Woong Choi contributed equally to this work.*)

REFERENCES

- [1] P. You, R. Instrum, and L. Parnes, "Benign paroxysmal positional vertigo," *Laryngoscope Investig Otolaryngol*, vol. 4, no. 1, pp. 116–123, 2018.
- [2] N. Bhattacharyya, S. P. Gubbels, S. R. Schwartz, J. A. Edlow, H. El-Kashlan, T. Fife, J. M. Holmberg, K. Mahoney, D. B. Hollingsworth, R. Roberts, M. D. Seidman, R. W. P. Steiner, B. T. Do, C. C. J. Voelker, R. W. Waguespack, and M. D. Corrigan, "Clinical practice guideline: Benign paroxysmal positional vertigo (update)," *Otolaryngology-Head Neck Surgery*, vol. 156, pp. 1–47, Mar. 2017.
- [3] S. Yetiser and D. Ince, "Demographic analysis of benign paroxysmal positional vertigo as a common public health problem," *Ann. Med. Health Sci. Res.*, vol. 5, no. 1, pp. 50–53, 2015.
- [4] M. Riga, A. Bibas, J. Xenellis, and S. Korres, "Inner ear disease and benign paroxysmal positional vertigo: A critical review of incidence, clinical characteristics, and management," *Int. J. Otolaryngology*, vol. 2011, Oct. 2011, Art. no. 709469.
- [5] Institute for Quality InformedHealth org Internet Cologne and Efficiency in Health Care (IQWiG). (2020). *Benign paroxysmal positional vertigo: Overview*. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK556947/>
- [6] D.-H. Liu, C.-H. Kuo, C.-T. Wang, C.-C. Chiu, T.-J. Chen, D.-K. Hwang, and C.-L. Kao, "Age-related increases in benign paroxysmal positional vertigo are reversed in women taking estrogen replacement therapy: A population-based study in Taiwan," *Frontiers Aging Neurosci.*, vol. 9, p. 404, Dec. 2017.
- [7] D. Balatsouras, G. Koukoutsis, A. Fassolis, A. Moukos, and A. Aspris, "Benign paroxysmal positional vertigo in the elderly: Current insights," *Clin. Intervent. Aging*, vol. 13, pp. 2251–2266, Nov. 2018.
- [8] H. Byun, J. H. Chung, S. H. Lee, C. W. Park, E. M. Kim, and I. Kim, "Increased risk of benign paroxysmal positional vertigo in osteoporosis: A nationwide population-based cohort study," *Sci. Rep.*, vol. 9, no. 1, pp. 2045–2322, Dec. 2019.
- [9] H. Jeremy, "Benign paroxysmal positional vertigo (BPPV): History, pathophysiology, office treatment and future directions," *Int. J. Otolaryngology*, vol. 2011, Jul. 2011, pp. 1687–9201.
- [10] H. Tang and W. Li, "Advances in the diagnosis and treatment of benign paroxysmal positional vertigo," *Experim. Therapeutic Med.*, vol. 14, no. 3, pp. 2424–2430, Sep. 2017.
- [11] K. A. Kerber and R. W. Baloh, "The evaluation of a patient with dizziness," *Neurol., Clin. Pract.*, vol. 1, no. 1, pp. 24–33, Dec. 2011.
- [12] D. G. Balatsouras, G. Koukoutsis, P. Ganelis, G. S. Korres, and A. Kaberos, "Diagnosis of single- or multiple-canal benign paroxysmal positional vertigo according to the type of nystagmus," *Int. J. Otolaryngology*, vol. 2011, pp. 1687–9201, Oct. 2011.
- [13] G. Lee, S.-G. Lee, H.-S. Park, B. J. Kim, S.-J. Choi, and J. W. Choi, "Clinical characteristics and associated factors of canal switch in benign paroxysmal positional vertigo," *J. Vestibular Res.*, vol. 29, no. 5, pp. 253–260, Nov. 2019.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.
- [15] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds. Long Beach, CA, USA, 2019, pp. 6105–6114. [Online]. Available: <http://proceedings.mlr.press/v97/tan19a.html>
- [16] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, "Self-training with noisy Student improves ImageNet classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10687–10698.
- [17] J. Lei, L. Yu, M. Bansal, and T. Berg, "TVQA: Localized, compositional video question answering," in *Proc. Conf. Empirical Methods Natural Lang. Process*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 1369–1379. [Online]. Available: <https://www.aclweb.org/anthology/D18-1167>
- [18] J. Kim, M. Ma, T. Pham, K. Kim, and C. D. Yoo, "Modality shifting attention network for multi-modal video question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10106–10115.
- [19] J. Donahue and K. Simonyan, "Large scale adversarial representation learning," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, 2019, pp. 10542–10552. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/file/18cdf49ea54eec029238fcc95f76ce41-Paper.pdf>
- [20] T. Vu, H. Jang, T. X. Pham, and C. Yoo, "Cascade RPN: Delving into high-quality region proposal network with adaptive convolution," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, 2019, pp. 1432–1442. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/file/d554f7bb7be44a7267068a7df88ddd20-Paper.pdf>
- [21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://www.aclweb.org/anthology/N19-1423>
- [22] R. Poplin, A. V. Varadarajan, K. Blumer, Y. Liu, M. V. McConnell, G. S. Corrado, L. Peng, and D. R. Webster, "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning," *Nature Biomed. Eng.*, vol. 2, no. 3, pp. 158–164, 2018.
- [23] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, Feb. 2017.
- [24] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

- [25] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, 2014, pp. 568–576. [Online]. Available: <https://proceedings.neurips.cc/paper/2014/file/00ec53c4682d36f5c4359f4ae7bd7ba1-Paper.pdf2>
- [26] B. Singh, T. K. Marks, M. Jones, O. Tuzel, and M. Shao, "A multi-stream Bi-directional recurrent neural network for fine-grained action detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1961–1970.
- [27] A. Ramaswamy, K. Seemakurthy, J. Gubbi, and B. Purushothaman, "Spatio-temporal action detection and localization using a hierarchical LSTM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 764–765.
- [28] A. B. Slama, H. Sahli, A. Mouelhi, J. Marrakchi, S. Boukriba, H. Trabelsi, and M. Sayadi, "Hybrid clustering system using nystagmus parameters discrimination for vestibular disorder diagnosis," *J. X-Ray Sci. Technol.*, vol. 28, no. 5, pp. 923–938, Sep. 2020.
- [29] A. B. Slama, A. Mouelhi, H. Sahli, S. Manoubi, R. Lahiani, M. B. Salah, H. Trabelsi, and M. Sayadi, "A new neural network method for peripheral vestibular disorder recognition using VNG parameter optimisation," *Int. J. Biomed. Eng. Technol.*, vol. 27, no. 4, pp. 321–336, 2018.
- [30] A. B. Slama, A. Mouelhi, H. Sahli, A. Zeraii, J. Marrakchi, and H. Trabelsi, "A deep convolutional neural network for automated vestibular disorder classification using VNG analysis," *Comput. Methods Biomech. Biomed. Eng., Imag. Visualizat.*, vol. 8, no. 3, pp. 334–342, May 2020, doi: 10.1080/21681163.2019.1699165.
- [31] E.-C. Lim, J. H. Park, H. J. Jeon, H.-J. Kim, H.-J. Lee, C.-G. Song, and S. K. Hong, "Developing a diagnostic decision support system for benign paroxysmal positional vertigo using a deep-learning model," *J. Clin. Med.*, vol. 8, no. 5, p. 633, May 2019.
- [32] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.
- [33] L. Wang, P. Koniusz, and D. Huynh, "Hallucinating IDT descriptors and 3D optical flow features for action recognition with CNNs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8697–8707.
- [34] K. B. Kwan-Loo, J. C. Ortiz-Bayliss, S. E. Conant-Pablos, H. Terashima-Marin, and P. Rad, "Detection of violent behavior using neural networks and pose estimation," *IEEE Access*, vol. 10, pp. 86339–86352, 2022.
- [35] J. D. Domingo, J. Gomez-Garcia-Bermejo, and E. Zalama, "Improving human activity recognition integrating LSTM with different data sources: Features, object detection and skeleton tracking," *IEEE Access*, vol. 10, pp. 68213–68230, 2022.
- [36] L. Lu, C. Zhang, K. Cao, T. Deng, and Q. Yang, "A multichannel CNN-GRU model for human activity recognition," *IEEE Access*, vol. 10, pp. 66797–66810, 2022.
- [37] J. Cha, M. Saqlain, D. Kim, S. Lee, S. Lee, and S. Baek, "Learning 3D skeletal representation from transformer for action recognition," *IEEE Access*, vol. 10, pp. 67541–67550, 2022.
- [38] J. Park, W.-S. Lim, D.-W. Kim, and J. Lee, "Multitemporal sampling module for real-time human activity recognition," *IEEE Access*, vol. 10, pp. 54507–54515, 2022.
- [39] Q. Zeng, M. O. Tezcan, and J. Konrad, "Dynamic equilibrium module for action recognition," *IEEE Access*, vol. 9, pp. 168015–168025, 2021.
- [40] H. Chen, M. Li, L. Jing, and Z. Cheng, "Lightweight long and short-range spatial-temporal graph convolutional network for skeleton-based action recognition," *IEEE Access*, vol. 9, pp. 161374–161382, 2021.
- [41] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, and J. Huang, "End-to-end learning of motion representation for video understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6016–6025.
- [42] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
- [43] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.
- [44] S. T. Moore, I. S. Curthoys, and S. G. McCoy, "VTM—An image-processing system for measuring ocular torsion," *Comput. Methods Programs Biomed.*, vol. 35, no. 3, pp. 219–230, Jul. 1991. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/016926079190124C>
- [45] R. O. Duda and R. E. Hart, "Use of the Hough transformation to detect lines and curves in pictures," *Commun. ACM*, vol. 15, no. 1, pp. 11–15, Jan. 1972.
- [46] T. Santini, W. Fuhl, and E. Kasneci, "PuRe: Robust pupil detection for real-time pervasive eye tracking," *Comput. Vis. Image Understand.*, vol. 170, pp. 40–50, May 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1077314218300146>
- [47] F. Vera-Olmos, E. Pardo, H. Melero, and N. Malpica, "DeepEye: Deep convolutional network for pupil detection in real environments," *Integr. Comput.-Aided Eng.*, vol. 26, no. 1, pp. 85–95, 2019.
- [48] S. Eivazi, T. Santini, A. Keshavarzi, T. Kubler, and A. Mazzei, "Improving real-time CNN-based pupil detection through domain-specific data augmentation," in *Proc. 11th ACM Symp. Eye Tracking Res. Appl.* New York, NY, USA: Association for Computing Machinery, Jun. 2019, pp. 1–6, doi: 10.1145/3314111.3319914.
- [49] M. Von Brevern, P. Bertholon, T. Brandt, T. Fife, T. Imai, D. Nuti, and D. Newman-Toker, "Benign paroxysmal positional vertigo: Diagnostic criteria," *J. Vestibular Res.*, vol. 25, nos. 3–4, pp. 105–117, 2015.
- [50] A. B. Slama, A. Mouelhi, H. Sahli, S. Manoubi, M. B. Salah, M. Sayadi, H. Trabelsi, and F. Fnaiech, "A novel automatic diagnostic approach based on nystagmus feature selection and neural network classification," in *Proc. 42nd Annu. Conf. IEEE Ind. Electron. Soc. (IECON)*, Oct. 2016, pp. 5165–5170.
- [51] Y.-H. Yiu, M. Aboulatta, T. Raiser, L. Oprey, V. L. Flanagan, P. Z. Eulenburg, and S.-A. Ahmadi, "DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning," *J. Neurosci. Methods*, vol. 324, Aug. 2019, Art. no. 108307. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0165027019301578>
- [52] A. M. Bronstein, "Vestibular reflexes and positional manoeuvres," *J. Neurol., Neurosurgery Psychiatry*, vol. 74, no. 3, pp. 289–293, Mar. 2003. [Online]. Available: <https://jnnp.bmj.com/content/74/3/289>
- [53] H. Touvron, A. Vedaldi, M. Douze, and H. Jegou, "Fixing the train-test resolution discrepancy," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, 2019, pp. 8252–8262. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/file/d03a857a23b5285736c4d55e0bb067c8-Paper.pdf>
- [54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Y. Bengio and Y. LeCun, Eds. San Diego, CA, USA, 2015, pp. 1–15.
- [55] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul. 1998.
- [56] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.
- [57] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.
- [58] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008. [Online]. Available: <http://jmlr.org/papers/v9/vandermaaten08a.html>
- [59] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [60] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–22.



TRUNG XUAN PHAM (Student Member, IEEE) received the B.S. degree from the School of Electronics and Telecommunications (SET), Hanoi University of Science and Technology (HUST), in 2014. He is currently pursuing the Ph.D. degree with KAIST under the supervision of Prof. Chang D. Yoo. His doctoral research interests include speech processing, self-supervised learning, and computer vision.



JIN WOONG CHOI (Senior Member, IEEE) received the Ph.D. degree from the Department of Otorhinolaryngology-Head and Neck Surgery, School of Medicine, Chungnam National University, in 2015. He is an Associate Professor with the Department of Otorhinolaryngology-Head and Neck Surgery, School of Medicine, Chungnam National University. His interests include otology and neurotology.



SULTAN RIZKY MADJID (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering with a double major at mechanical engineering from KAIST, in 2021, where he is currently pursuing the M.S. degree. He has been a Machine Learning Researcher with KAIST, since 2021. His research interests include model compression, sparse representations in deep learning, and continual learning.



RUSTY JOHN LLOYD MINA (Student Member, IEEE) received the B.S. degree in electronics and communications engineering from the University of the Philippines Diliman, in 2018, and the M.S. degree from KAIST, in 2021. He was a Machine Learning Researcher at KAIST, from 2019 to 2021. His research interests include ML art, computer vision, and continual learning.



THANH NGUYEN (Graduate Student Member, IEEE) received the B.S. degree in electronic and automation engineering from the Ho Chi Minh City University of Science and Technology, in 2015. He is currently pursuing the M.Sc. and Ph.D. degrees with the Korea Advanced Institute of Science and Technology. His research interests include machine learning, deep learning, and reinforcement learning.



CHANG D. YOO (Senior Member, IEEE) received the B.S. degree in engineering and applied science from the California Institute of Technology, the M.S. degree in electrical engineering from Cornell University, and the Ph.D. degree in electrical engineering from MIT. He is the Director of the Video Turing Test Research Center and the AI Fairness Research Center. His current research interests include machine learning, signal processing, computer vision, and audio processing. He is a member of Tau Beta Pi and Sigma Xi.

...