Biomedical Signal Processing and Control 68 (2021) 102616

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Deep learning based torsional nystagmus detection for dizziness and vertigo diagnosis

Wanlu Zhang [a,b], Haiyan Wu [c,*], Yang Liu [a,d], Shuai Zheng [a,b], Zhizhe Liu [a,b], Youru Li [a,b], Yao Zhao [a,b], Zhenfeng Zhu [a,b]

[a] *Institute of Information Science, Beijing Jiaotong University, Beijing, China*
[b] *Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing, China*
[c] *Department of Otolaryngology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China*
[d] *School of Information Science and Engineering, Hebei North University, Zhangjiakou, China*

ARTICLE INFO

Keywords:
Torsional nystagmus detection
Deep learning
Video condensation
Pupil calibration
Optical field flow

MSC:
00-01
99-00

ABSTRACT

Dizziness and vertigo are common clinical symptoms and typical complaints of many vestibular diseases. In the bedside examination of dizziness and vertigo, nystagmus is the most sensitive and specific sign of vestibular lesions. The measurement of nystagmus pattern by infrared video goggle collected in clinic can provide a valuable diagnostic information for dizziness and vertigo. This paper mainly studies the automatic detection of torsional BPPV nystagmus based on deep learning, thus assisting clinicians diagnose dizziness and vertigo conveniently. In order to eliminate the invalid frames from the blinking of patients under observation, a convolutional neural network(ConvNet) based eye movement video condensation approach is proposed. When calibrating the moving pupil in the captured frame sequence, the Hough transform and trajectory tracking based on template matching are well combined to improve the robustness to eyelash occlusion and pupil deformation. In addition, the optical flow field of moving eyeball is exploited to characterize the torsion motion of torsional nystagmus, based on which a Torsion-aware Bi-Stream Identification Network (TBSIN) model is proposed. Furthermore, through label-error correction based on temporal consistency, we can merge multiple continuous torsional frames into torsional nystagmus segments for clinical diagnosis. Experiments are conducted on a clinically collected torsional nystagmus video dataset and promising experimental results show the effectiveness of the proposed approach. In particular, we achieve 85.73% and 81.00% in view of Accuracy and F1 measurements for frame-level identification, as well as IOU performance 67.45% for final torsional nystagmus segment localization.

1. Introduction

Vestibular system is an important part of human balance system. As the main organ for human body to perceive the changes of body position and environment, it plays a key role in the sense of balance and maintaining stable vision and posture. In clinic, the most common symptoms of vestibular diseases are dizziness and vertigo (abbr. by DaV), which are also the primary complaints. According to some statistics, its incidence rate reaches 20–30% in general population [1]. In addition, 40% of adults have obvious dizziness symptoms [2], while more than 50% [3] of the elderly suffer from similar symptoms. Given the noteworthy prevalence of DaV, it inflicts a considerable personal and socio-economic burden. Therefore, the research on DaV has been paid more and more attention, especially on the diagnosis of DaV.

No matter for what kind of disease diagnosis, medical history is indispensable. The detailed and complete medical history is always the first choice of disease diagnosis. In addition to the medical history record for the diagnosis of DaV, the patient's physical signs are another important basis for the diagnosis of DaV. A large number of studies have shown that there exists a close coupling relationship between abnormal eye movements and vestibular disorders [4]. Therefore, as an involuntary, rapid, and rhythmic movement of the eyeball, nystagmus has been popularly regarded as the most obvious and important sign in various vestibular disorders. According to the direction of eye movement, nystagmus patterns can be divided into horizontal, vertical, diagonal, and torsional. In the actual clinical signs of patients, these nystagmus

* Corresponding author.
E-mail address: ywjwhy@sina.com (H. Wu).

https://doi.org/10.1016/j.bspc.2021.102616
Received 5 September 2020; Received in revised form 30 March 2021; Accepted 4 April 2021
Available online 21 April 2021

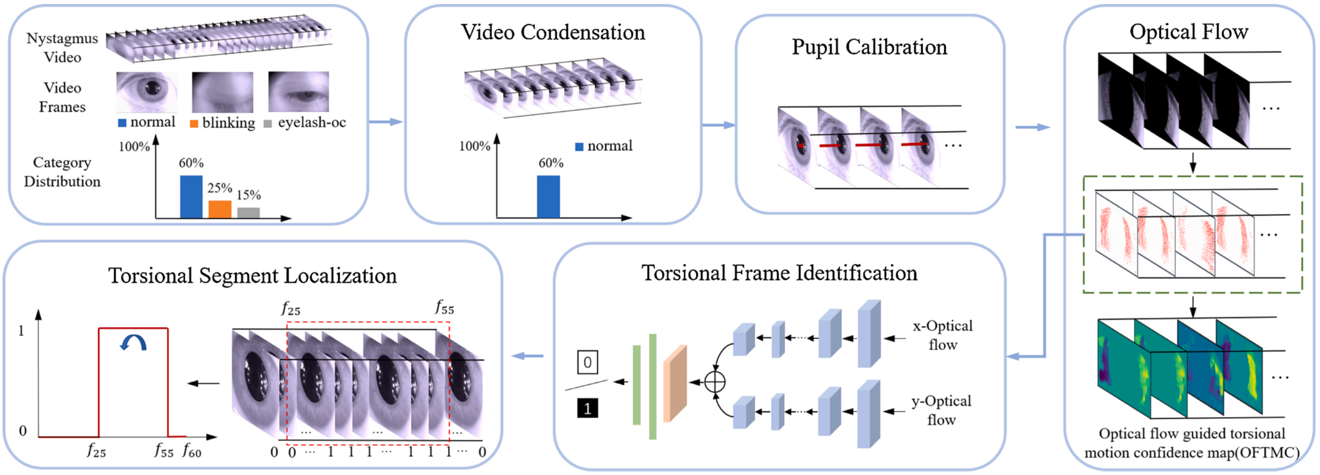


Fig. 1. Deep learning based framework for torsional nystagmus detection, which mainly consists of video condensation, pupil calibration, dense optical flow estimation, deep learning based torsional frame identification and torsional segment localization.

patterns are usually presented in a mixed way, such as "horizontal + torsional" or "vertical + torsional".

Different from the traditional observation of patients' physical signs by naked eye, video nystagmograph (VNG) [5,6] has been widely used in the clinical diagnosis of dizziness and vertigo. With VNG, the pupil movement information of patients can be captured to output sawtooth time series signal from nystagmus. Through the analysis of nystagmus signal, we can get some of the characteristics of nystagmus, including nystagmus trajectory, velocity, waveform, frequency, amplitude, temporal profile [7], and so on, so as to assist doctors in the diagnosis of dizziness and vertigo. However, it should be noted that the time series signals of nystagmus obtained through VNG only reflect the horizontal and vertical translation information of the eyeball, and lack the ability to capture the torsion motion of the eyeball.

However, for benign paroxysmal positional vertigo (BPPV), the most frequent vestibular disorder [8–11], torsional nystagmus is the most diagnostic signs associated with it. As a peripheral vestibular disease, BPPV refers to the movement of a head to a specific location, which can induce a short paroxysmal vertigo, accompanied by nystagmus and autonomic symptoms. The incidence rate is as high as 17–22% in patients with vertigo. According to its diagnostic positional maneuvers and a canal-specific nystagmus characteristics, BPPV is the most commonly clinically encountered as one of two variants: BPPV of the posterior semicircular canal (pc-BPPV) [12] or BPPV of the horizontal semicircular canal (hc-BPPV). Among them, pc-BPPV is by far the most common variant, accounting for 85–95% of the cases [13]. Unlike the hc-BPPV whose nystagmus was mainly horizontal, the pc-BPPV nystagmus often presents more complex patterns, with torsional and upbeating vertical nystagmus. For this reason, the above characteristic parameters provided by the traditional instrument like VNG cannot give doctors objective indicators to make auxiliary judgments, which will bring great challenges to the diagnosis of some non-specialist doctors. In addition, even for doctors with rich clinical experience, it is also not trivial to judge torsional nystagmus quickly and accurately by visual inspection.

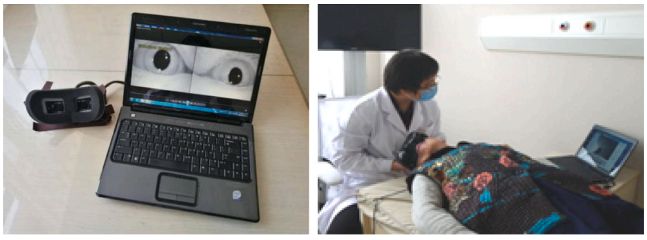


Fig. 2. Portable eye movement video acquisition equipment.

In recent years, we have witnessed the advance of deep learning, which has greatly promoted the development of computer vision [14–17], natural language processing [18–20] and other artificial intelligence technologies. In the aspect of medical intelligent auxiliary diagnosis, deep learning has also been successfully applied, such as the auxiliary diagnosis of benign and malignant of chest nodule CT images [21,22], as well as the automatic analysis of skin disease images [23,24], fundus images of eye diseases [25,26], pathological sections of malignant tumors [27,28], etc. In order to reduce the difficulty of clinical diagnosis of DaV and improve the efficiency of diagnosis, we mainly focus on developing a deep learning based framework for torsional nystagmus detection in this paper. Here, the detection of torsional nystagmus refers to locate the torsional segments consisting of multiple temporal continuous torsional frames.

To the best of our knowledge, less effort has been devoted to the automatic detection of torsional nystagmus in previous work due to the difficulties in characterizing the extremely weak, imperceptible torsional motion. The following points highlight several contributions of the paper:

- This paper proposes a deep learning based framework to realize the automatic detection of torsional nystagmus, which can assist clinicians diagnose DaV conveniently.
- To eliminate the invalid frames from the blinking of patients under observation and lighten the workload of doctors in video browsing, we propose an efficient convolutional neural network based approach for eye movement video condensation.
- For the captured sequence of frames containing eyeball, we combine the Hough transform and template matching based trajectory tracking to calibrate the moving eyeball, facilitating the robustness to eyelash occlusion and pupil deformation.
- To characterize the torsion pattern of torsional nystagmus, the dense optical flow field is well exploited to establish the confidence map of torsion motion intensity, on which the Torsion-aware Bi-stream Identification Network (TBSIN) model is proposed to identify the torsional frames.
- We collect a benchmark dataset about pc-BPPV with manual annotation, which can serve as a good benchmark for research community.

The rest of this paper is organized as follows: Section 2 gives an illustration on the proposed deep learning based framework for detection of torsional nystagmus. Section 3 presents the methods of nystagmus video condensation and calibrating. In Section 3, we show

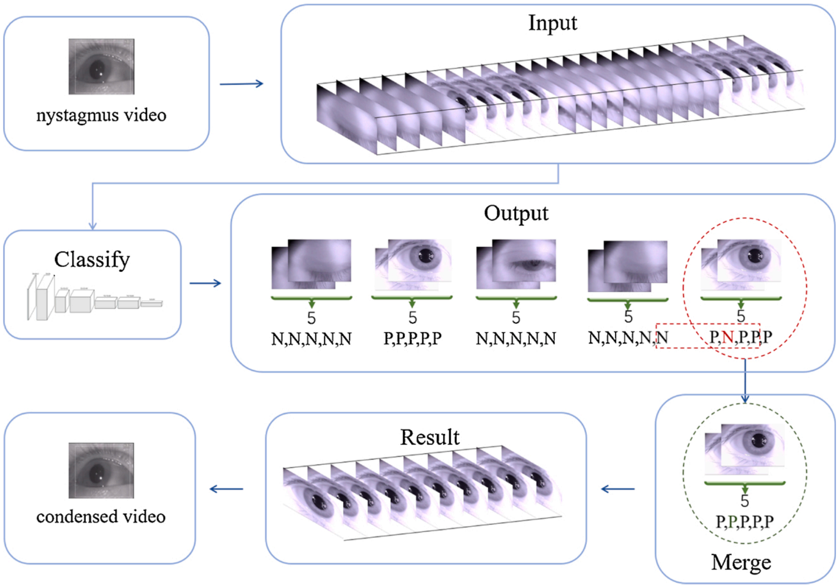


Fig. 3. Illustration of eye movement video condensation.

the details about torsional nystagmus detection based on deep learning. Section 5 demonstrates the experimental results and analysis, and Section 4 concludes this paper.

2. Framework

The overall framework of the proposed deep learning based torsional nystagmus detection is shown in Fig. 1. Different from the commonly used videonystagmoscopy to collect eyeball movement video, we have developed a portable infrared video goggle to capture the eyeball movement as shown in Fig. 2. In this way, unlike the VNG installed in the inspection room, it is more convenient to carry out some bedside inspections that are not limited by time and place.

For the captured eye movement video, the pipeline processing mainly includes video condensation, pupil calibration, dense optical flow estimation, deep learning based torsional frame identification, and torsional segment localization.

- *Video condensation.* The purpose of video condensation is to eliminate those invalid frames with eye close, eyelash occlusion, etc.
- *Pupil calibration.* For torsional nystagmus, the corresponding eye movement is usually composed of compound movement. Therefore, it is necessary to calibrate pupil center before estimating torsion intensity through optical flow field.
- *Deep learning based torsional frame identification.* It should also be pointed out that, the torsional frame identification here refers to the frame-level torsion recognition.
- *Torsional segment localization.* On the basis of torsional frame identification, the torsional segment localization means the merging of multiple temporal-continuous torsional frames into independent torsional segments.

3. Condensation and calibrating of eye movement video

3.1. CNN-based eye movement video condensation

In bedside vestibular function examination based on eye movement video, due to the uncontrollable acquisition equipment, environment and individual patient differences, there are inevitably a large number of invalid video frames of the acquired video, such as those existing in blinking, eyelash occlusion. The purpose of video condensation is to eliminate these invalid frames, which will be beneficial to avoid the interference of them to the judgment of torsional nystagmus. On the other hand, it can help clinician to improve efficiency in taking retrospective look over the captured video. In addition, it will also facilitate savings in video storage.

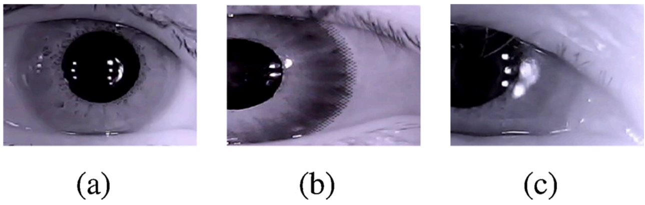


Fig. 4. The cases of spot noise. (a) spot noise inside pupil; (b) spot noise on junction location; (c) spot noise outside pupil.

In Fig. 3, we show the proposed convolutionary neural network (CNN) [29,30] based video condensation. In practice, the video consideration can be taken as a task of binary classification, for which an end-to-end CNN model ConvNets is trained on a labeled training dataset. Hence, for each frame of the input video, it can be classified as *P*-frame (i.e., valid frame) or *N*-frame (i.e., invalid frame). Here, the *P*-frame also denotes positive sample in model training, while *N*-frame denoting negative sample. Considering the possible misclassification of ConvNets, the temporal context-consistency is considered for error correction. Specifically, we set a sliding window of size $2s+1$ (s=2 in our case) on each frame f_i, and then calculate the number C_i of frames with the same label as f_i in the window. Once C_i is less than a threshold T_c, its label will be corrected. Finally, a condensed eye movement video can be obtained by putting all video frames labeled as *P* together.

3.2. Pupil calibration of eye movement video

For torsional nystagmus, the eye movement is often very complex as we mentioned above, including both torsional motion and translational motion in *x*-direction and *y*-direction. To give an accurate estimation on torsion motion, the eyeballs from different frames need to be calibrated first to eliminate translational motion. Since the pupil can be approximated to a circle, a straightforward way is to use the center of the pupil as a calibration reference point.

One point we should note in the process of eye movement video acquisition, the reflection of light source will cause spot noise on pupil in

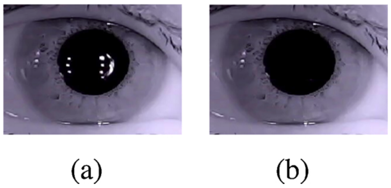


Fig. 5. Elimination of spot noise by using morphological pre-processing. (a) pupil with spot noise; (b) pupil with spot noise eliminated.

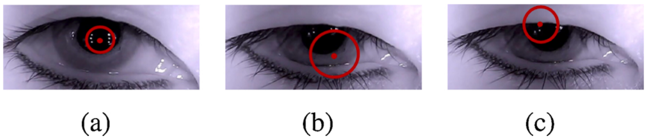


Fig. 6. Poorly positioned results by CHT.

many cases as shown in Fig. 4. The presence of spot noise will adversely affect pupil calibration. Hence, a simple but effective morphological [31] pre-processing was performed before pupil calibration. Fig. 5(b) shows one of results with morphological pre-processing.

As a parameterized model, Circular Hough Transform (CHT) has been widely used for localizing the circle center [32,33]. However, due to the influence of eyelash interference and pupil deformation, CHT does not always perform well in pupil center localization. Fig. 6 shows several poorly positioned examples by CHT. In the case of the calibration of eye movement video, it is essentially a joint spatio-temporal positioning problem. In addition to the spatio-positioning of pupil in each frame, the potential temporal correlations among consecutive frames should also be exploited to make a good alignment of them. In particular, to improve the robustness to pupil deformation, the pupil light spot, and occlusion from eyelash and eyelid, we propose a spatio-temporal consistent moving eyeball calibrating approach by combining the CHT and template matching based trajectory tracking.

We illustrate the proposed procedure of pupil calibration in Fig. 7. As we can see, the CHT is applied to the binarized image instead of the original input, which can be helpful of avoiding the influence of noisy edges. To deal with the multiple candidate circles obtained by CHT, the proportion of black pixels in Hough circle is used to find the best matching. Once we find the initial calibration reference point, i.e., the center of the circle with the maximum radius (the green circle), the template matching based trajectory tracking is carried out to keep temporal consistency of calibration. For the detailed implementation of eye movement video, we summarize it in Algorithm 1.

Algorithm 1. Pupil calibration of eye movement video

```
Input:
$f[i]$, $i = 1, \ldots, N_f$: sequence of frames
$P = TM(A, B)$: template matching to find the optimum matching point $P$ of $A$ in $B$.
$Thr$: threshold for template updating
Output:
$f_i \in \mathbb{R}^{h\times w}$, $i = 1, \ldots, N_f$: sequence of calibrated frames
for $i = 1, \ldots, N_f$ do
    Given the input $f[i]$, obtain the morphologically pre-processed $\mathrm{Frame}[i]$.
    $\left\{\left(C_i^j, P_i^j\right)\right\}_{j=1,\cdots,N_c} = \mathrm{CHT}(\mathrm{Frame}[i])$. // Obtain $N_c$ multiple candidate circles
      and corresponding centers for pupil fitting by Circular Hough Transform on binarized $\mathrm{Frame}[i]$.
    $C_i^{\mathrm{jMax}} \rightarrow C_i^H$, $P_i^{\mathrm{jMax}} \rightarrow P_i^H$. // As the best fitting of pupil in frame[$i$], $C_i^{\mathrm{jMax}}$ is with
      the largest percentage of black pixels.
end for
$C_{\mathrm{iMax}}^H \rightarrow T_b$, $C_{\mathrm{iMax}}^H \rightarrow T_f$, $P_{\mathrm{iMax}}^H \rightarrow P_{\mathrm{iMax}}^T$. // Initializes the template using the circle
  $C_{\mathrm{iMax}}^H$ with the maximum radius.
for $m = iMax - 1, \ldots, 1$, $n = iMax + 1, \ldots, N_f$ do
    $\mathrm{TM}(T_b, \mathrm{Frame}[m]) \rightarrow P_m^T$.
    $\mathrm{TM}(T_f, \mathrm{Frame}[n]) \rightarrow P_n^T$.
    if $||P_b - C_m^H||^2 \geq \mathrm{Thr}$then
        $C_m^H \rightarrow T_b$, $P_m^H \rightarrow P_m^T$.
    endif
    if $||P_f - C_n^H||^2 \geq \mathrm{Thr}$ then
        $C_n^H \rightarrow T_f$, $P_n^H \rightarrow P_n^T$.
    endif
    Centered on $P_m^T$ and $P_n^T$, respectively, cropping sub-images $f_m$ and $f_n$ with size
      $h$ by $w$.
end for
return $f_i$, $i = 1, \ldots, N_f$.
```

4. Torsional nystagmus detection based on deep convolutional network

4.1. Optical flow guided torsional motion confidence map

For the detection and segmentation of torsional nystagmus, the central issue is how to determine the torsion of the eyeball. In clinic, when doctors are watching the nystagmus videos, they usually focus on

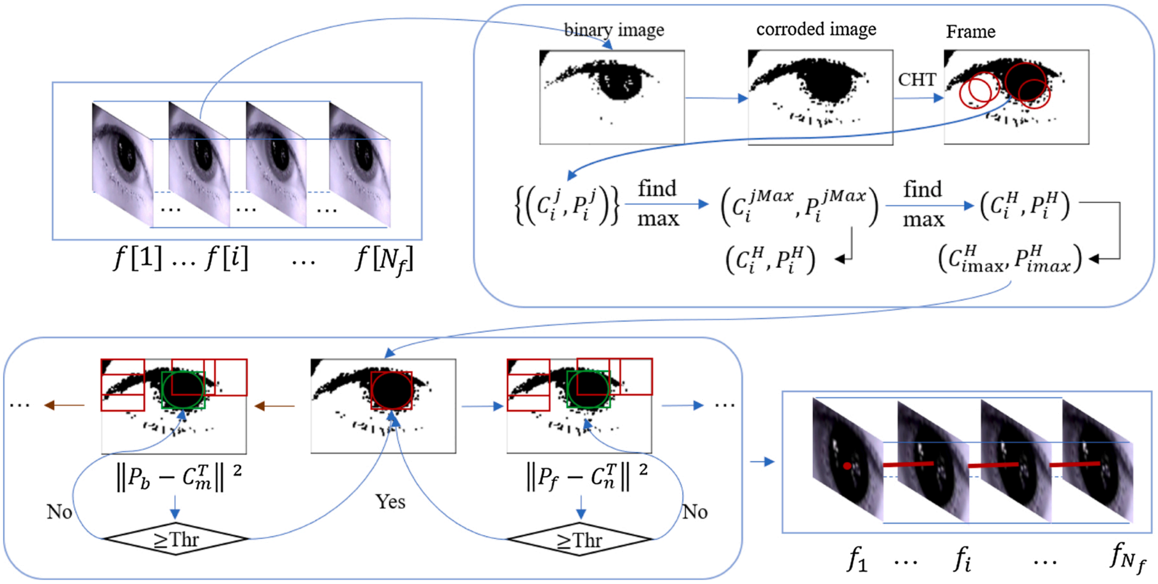

binary image
corroded image
Frame
CHT
$f[1] \ldots f[i] \quad \ldots \quad f[N_f]$
$\left\{\left(C_i^j, P_i^j\right)\right\}$
find max
$\left(C_i^{jMax}, P_i^{jMax}\right)$
find max
$\left(C_i^H, P_i^H\right)$
$\left(C_i^H, P_i^H\right)$
$\left(C_{imax}^H, P_{imax}^H\right)$
$\left\|P_b - C_m^T\right\|^2$
$\left\|P_f - C_n^T\right\|^2$
≥Thr
≥Thr
No
Yes
No
$f_1 \quad \cdots \quad f_i \quad \cdots \quad f_{N_f}$

Fig. 7. Pupil calibration based on circular hough transform and temporal template matching.

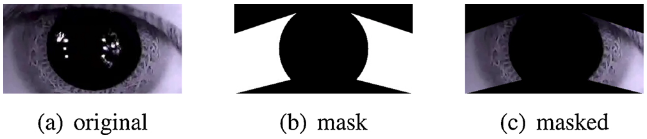

(a) original (b) mask (c) masked

Fig. 8. Binary masking for ROI extraction.

the visual changes in the area of the iris and sclera on the left and right parts of the pupil, where is just the region of interest (ROI) that can provide sufficient information for diagnose of torsional nystagmus. For this reason, a binary mask as shown in Fig. 8(b) is applied to keep only the ROI active for torsion estimation.

Optical flow exhibits the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and a scene, which can provide detailed two-dimensional motion vector information of the target from first frame to second [34–36]. Compared with sparse optical flow, dense optical flow is not just to select certain feature points in the image for matching, but to match the image pixel by pixel, calculate the offset of all points, and finally get the optical flow field [37,38].

For the eye movement video, there are no stable points of interest that can be extracted for calculating the sparse optical flow. To estimate the motion field of moving eyeball, instead of the sparse optical flow, the dense optical flow algorithm, also known as Gunnar Farneback method [39], is used in our case. Clinically, the difficulty in the diagnosis of torsional nystagmus mainly lies in the accompanying torsion motion is imperceptible. In order to make such weak torsion motion be apt to detect, the dense optical flow with intervals of s ($s = 2$ in our case) frames, not between the adjacent frames as usual, is calculated as shown in Fig. 9. In this way, it means we can actually make a more in-depth observation on the motion pattern of eyeball movement by enlarging the intensity of its movement.

In fact, it can be assumed that the optical flow field is equivalent to the motion field. Thus, in order to more intuitively reflect the movement trend and intensity of torsional nystagmus, we can establish an optical flow guided torsion motion confidence (OFTMC) map to give an intuitive characterization of torsion motion by using the obtained two-dimensional optical flow field. Let $\overrightarrow{\mathrm{F}}^m$ denote the optical flow field of the mth frame image f_m with:

$$\overrightarrow{\mathrm{F}^m}(i,j) = (M^m(i,j), \theta^m(i,j)) \tag{1}$$

where $M^m(i,j) = \sqrt{\Delta x^m(i,j)^2 + \Delta y^m(i,j)^2}$ is the intensity of optical flow, $\theta^m(i,j) = \arctan\left(\frac{\Delta y^m(i,j)}{\Delta x^m(i,j)}\right)$ is the motion vector angle, $\Delta x^m(i, j)$ and $\Delta y^m(i, j)$ represent the displacements in x-axis and y-axis of pixel $p(i, j)$, respectively. Hence, we can obtain the OFTMC map $C_r^m = \{C_r^m(i,j)\} \in \mathbb{R}^{h\times w}$ as follows:

$$C_r^m(i,j) = \mathrm{sgn}(\theta^m(i,j))\cdot M^m(i,j) \tag{2}$$

where $\mathrm{sgn}(\theta)$ is a sign function and given by:

$$\mathrm{sgn}(\theta) = \begin{cases} 1, 0 < \theta < \pi \\ -1, -\pi < \theta < 0 \end{cases} \tag{3}$$

From Eq. (2) we can find that the $M(i, j)$ is used to characterize the motion intensity of pixel $p(i, j)$, while the sign function $\mathrm{sgn}(\theta(i, j))$ on the motion vector angle $\theta(i, j)$ acts as a reflection of the torsion characteristics of pixel $p(i, j)$. To give a clearer explanation, the visualization of C_r is also shown in Fig. 9 by using pseudo color. The difference in color for the left and right regions of the iris and sclera in the eyeball explicitly shows complete opposite motion trend. This phenomena is very much in line with the clockwise right-handed torsional movement of eyeball (the upper pole of the eyes beating toward right) in frame f_i. In addition, several cases of OFTMC corresponding to different motion pasterns of eyeball movement are illustrated in Fig. 10. Compared with non-torsional frames in Fig. 10(a) and (b), the OFTMCs from Fig. 10(c) and (d) provide powerful clues to torsion motion. Particularly, taking the left-handed torsion (the upper pole of the eyes beating toward left) in 10 (c) as an example, the dominated blue color of pixels in the left region means a general downward movement trend, while the yellow color in the right region means a general upward motion trend. The opposite motion trend of the left and right regions constitutes the levogyration. Likewise, the OFTMC in Fig. 10(d) indicates a dextroclination of eyeball.

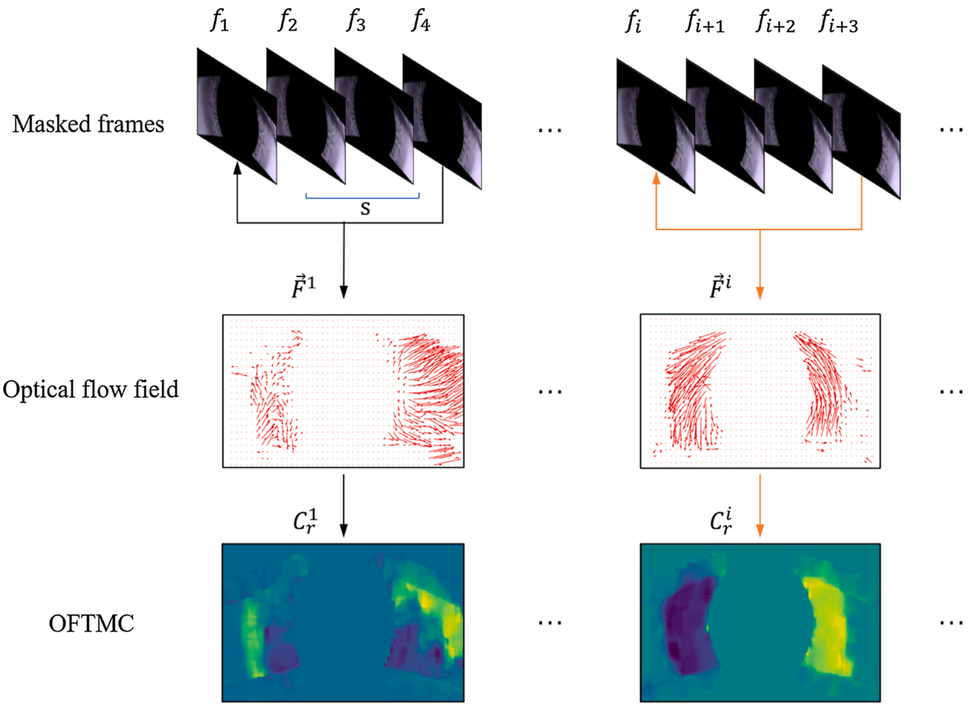
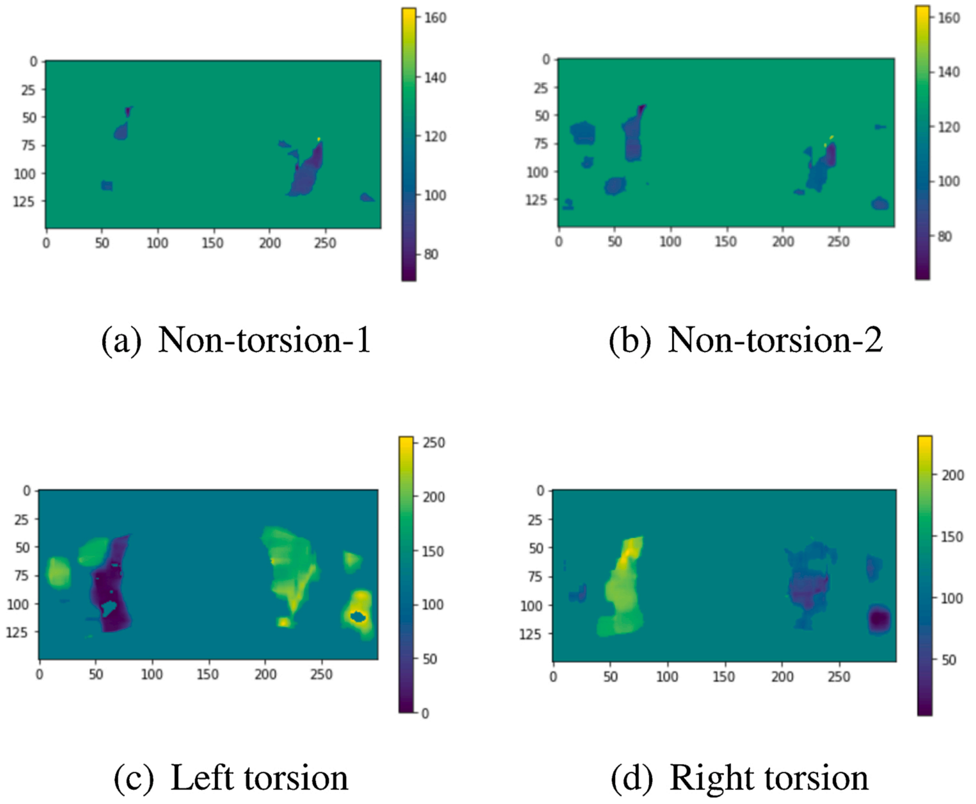


Fig. 9. Optical flow field guided torsion motion confidence map (OFTMC).



(a) Non-torsion-1 (b) Non-torsion-2

(c) Left torsion (d) Right torsion

Fig. 10. OFTMC of different motion patterns.

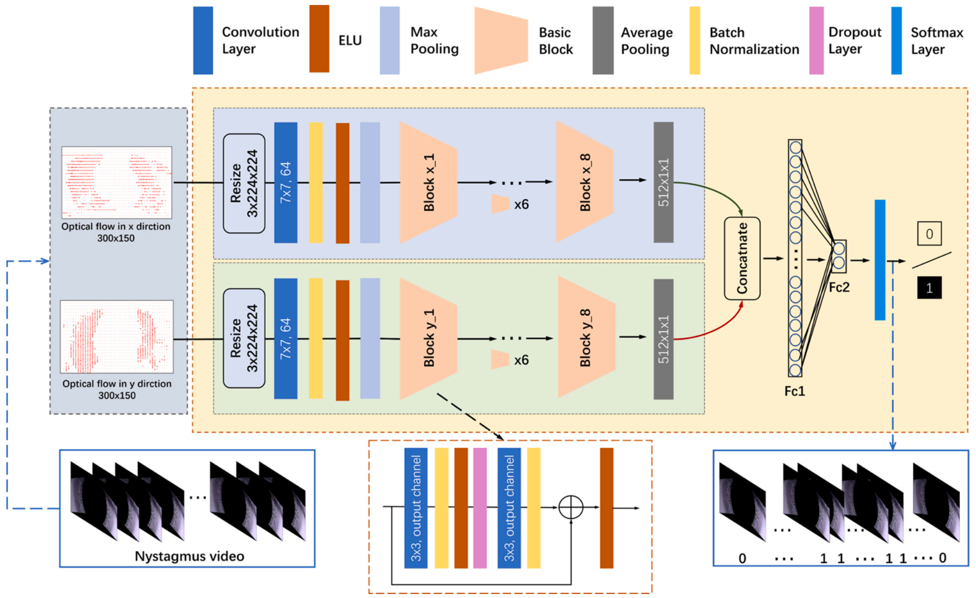


Fig. 11. Architecture of the proposed Torsion-aware Bi-Stream Identification Network (TBSIN) for torsional nystagmus video.

By establishing the OFTMC, the torsion motion and intensity of torsional nystagmus can be visually and intuitively observed.

4.2. Torsion-aware Bi-Stream Identification Network

The purpose of torsional nystagmus detection is to locate automatically those segments composed of continuous torsional frames in a nystagmus video. As mentioned above, for a torsional frame, its torsion motion pattern can be well revealed by OFTMC. In recent years, deep convolutional networks have exhibited a remarkable ability in image segmentation [40], object detection [15], image classification [17] and other vision-related tasks. To identify whether the input frame is torsional or not, a Torsion-aware Bi-stream Identification Network (TBSIN) model is proposed by exploiting the optical flow field associated with pupil movement.

Fig. 11 shows the architecture of the proposed TBSIN model. Instead of applying OFTMC directly as the input of the model, the optical flows in both *x*-direction and *y*-direction are input into the bi-stream network,

Table 1
Network architecture of ResNet18.

Layer name	Stack 1	Stack 2	Stack 3	Stack 4
Output size	$56 \times 56 \times 64$	$28 \times 28 \times 128$	$14 \times 14 \times 256$	$7 \times 7 \times 512$
18-layer	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$

thus making the identification model to be more aware of the torsional pattern. Considering the good performance of ResNet in feature extraction, the ResNet18 [17,41] network is used to serve as the backbone of TBSIN.

Specifically, we adopt the same network structure as ResNet18 for both x-subnetwork and y-subnetwork. Following a concatenation of two 512-d feature vectors from x-subnetwork and y-subnetwork, the network ends with two fully-connected layers (1024-way fc1 and 2-way fc2) and a softmax layer to output the predicted label '1' for torsional frame or '0' for non torsional frame. For each subnetwork, it consists of 1 root block and 4 stacks (Stack 1–4) with each stack being superimposed by two Basicblocks, about which the details are provided in Table 1. In addition, a Dropout layer is plugged into each Basicblock so as to avoid the overfitting in network training. It is worth noticing that since the input optical flows $\Delta x(i,j)$ and $\Delta y(i,j)$ at pixel $P(i,j)$ take either positive or negative value that corresponds to positive or negative direction, respectively, in x and y directions, thus we substitute the activation function Relu [42] in ResNet18 for Elu [43].

4.3. Label-error correction based on temporal label consistency

Although the torsional characteristics of the input frame can be well determined by TBSIN, there inevitably exist the cases of misclassification. In order to eliminate these errors, a label-error correction is proposed by exploiting label consistency in temporal.

Let $L = [l_i]_{i=1,\ldots N_f} \in \{0,1\}^{N_f}$ denote the set of labels output through TBSIN, where N_f is the number of frames. For each l_i, we put the center of a sliding windows W with size $2N_w + 1$ on $i - N_w$, i, and $i + N_w$ of L, respectively, to exploit the label consistency in temporal. To derive the judgement for label-error correction, we first count the number $c_{l_i}^j, j = \{1,$ 2, 3}, of labels in each window that are same as l_i. Then, we can make several individual judgements J_i^j by:

$$J_i^j = \begin{cases} \text{true}, & c_{l_i}^j \geq T, \\ \text{false}, & c_{l_i}^j < T \end{cases} \tag{4}$$

where T is the predefined threshold. Furthermore, by merging these individual judgements, we can get an ensemble judgement:

$$J_i = \begin{cases} \text{true}, & \text{Vote}(J_i^1, J_i^2, J_i^3) \geq 2, \\ \text{false}, & \text{else} \end{cases} \tag{5}$$

where $\text{Vote}(J_i^1, J_i^2, J_i^3)$ denotes the number of judgments that are true by the collaborative voting from J_i^1, J_i^2, and J_i^3. For the ith frame, once the obtained J_i is false, its label l_i will be changed to its opposite, i.e., $\widetilde{l_i} = 1 - l_i$; otherwise, it will remain unchanged. In Fig. 12, we give an illustration on label-error correction by label consistency in temporal. As we can see, the predicted label of the ith frame is clearly not correct. By exploiting the label consistency in temporal, the wrong label prediction can be effectively corrected.

Once the corrected sequence $\widetilde{L} = [\widetilde{l_i}]_{i=1,\ldots N_f} \in \{0,1\}^{N_f}$ of labels is obtained, we need to make a further step to locate the specific torsional nystagmus segments. To address this issue, a simple and heuristic way is adopted. Let's divide uniformly $\widetilde{L}$ into multiple blocks, each of which is with the interval of N_b frames ($N_b = 15$ in our case). For each block, if the number c_p of label '1' (torsional frame) is greater than the number c_n of label '0' (non-torsional frame), then all the labels in this block will be uniformly set to '1', otherwise to '0'. In this way, the boundary of the specific torsional nystagmus segments can be well localized. In essence, it can be seen as a block-level boundary determination of torsion segment.

5. Experimental results and analysis

5.1. Experiment setup

5.1.1. Datasets

The dataset used throughout this paper was collected in the Department of Otolaryngology, Peking Union Medical College Hospital. All inspection videos were captured through a portable infrared video goggle as shown in Fig. 2. For eye movement video condensation, we construct a dataset consisting of 4000 valid frames (positive samples)

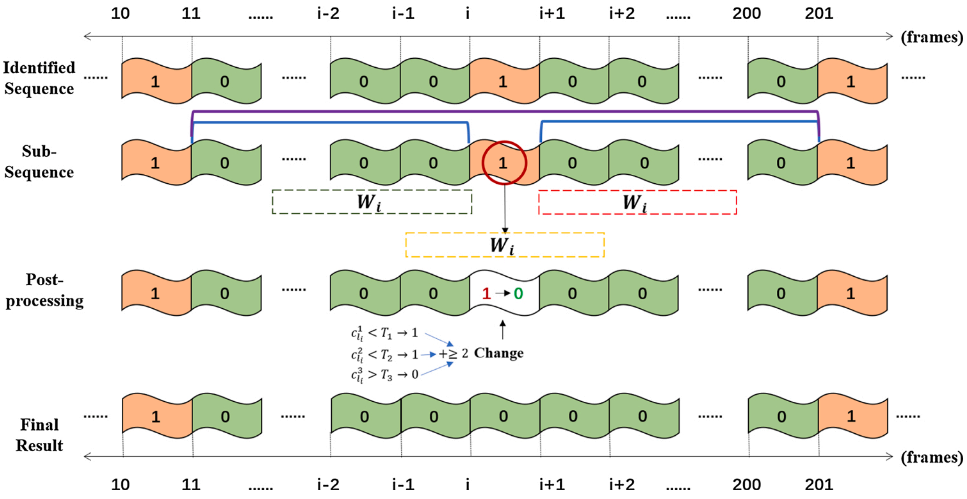


Fig. 12. Label-error correction based on temporal label consistency.

Table 2
The details of eye movement image datasets.

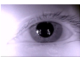
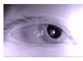
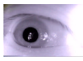
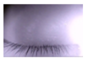
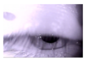
Category	Positive				Negative			
Feature	Pupil in the center	Pupil to the right	Pupil to the left	Pupil up	No pupil	Eyebrow	Eyelash	Pupil obscured
No. of frames	1000	1000	1000	1000	1000	1000	1000	1000
Reference image								

Table 3
The dataset for detection of torsional nystagmus.

Dataset	Torsional frames	Non-torsional frames
Training dataset	29,504	54,207
Testing dataset	7028	12,035

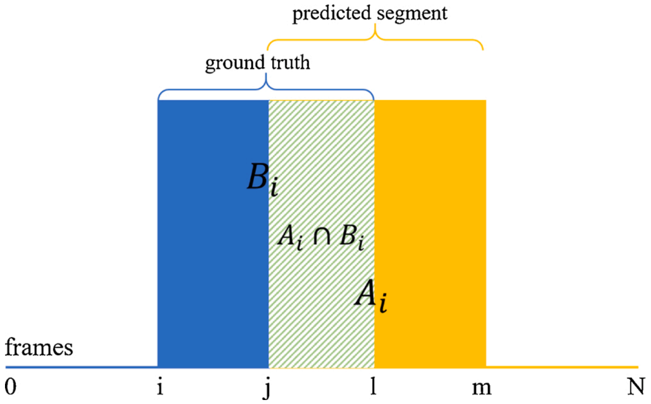


Fig. 13. The graphic explanation of *IoU*.

and 4000 invalid frames (negative samples), of which the details are given in Table 2. In particular, the invalid frames mainly include the cases of no pupil, eyebrow, eyelash, and pupil obscured. To train the condensation model, we use 80% of the total 8000 samples as training dataset and the remains as test dataset.

In addition, as a diagnostic positional test, the Dix-Hallpike maneuver was performed on all patients suspected of pc-BPPV. Totally, the collected dataset includes 77 pc-BPPV nystagmus videos, all of which are with torsional pattern, and each video contains at least one torsional nystagmus segment. The labelling of torsional nystagmus segments on each video is performed by a self-built nystagmus video annotation system. The detailed statistic about the dataset for torsional nystagmus detection are provided in Table 3.

5.1.2. Evaluation metrics

Performance evaluation measures of *Precision*, *Recall*, *F1*-score, and *Accuracy* were adopted to evaluate the performance of torsional frame identification of nystagmus video.

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{9}$$

where *TP*, *TN*, *FP*, and *FN* represent true positive, true negative, false positive, and false negative, respectively.

Table 4
The hyper-parameters of training ConvNets.

Stage	Hyper-parameters	Value
Structure	Input	224*224*3
	Convolution kernel	3x3
Training	Batch size	128
Adam	Learning rate	5.0×10^{-5}
	Weight decay	1.0×10^{-3}
Training	Epoch	20

Table 5
Performance comparisons of torsional frame identification.

Input	*Precision*	*Recall*	*Accuracy*	*F1*
x-optical flow	48.07	29.20	62.27	36.33
y-optical flow	62.16	66.19	72.68	64.11
OFTMC C_r	61.59	68.03	70.12	64.65
xy-optical flow	61.60	70.22	72.88	65.62

To give a quantitative performance evaluation on the detection of torsional segments of nystagmus videos, the following Intersection over Union (IoU) measurement as shown in Fig. 13 was used.

$$\text{IoU}_{\text{avg}} = \frac{\sum_{i=1}^{N_t} \frac{1}{N_i^g} \sum_{j=1}^{N_i^g} \sum_{k=1}^{N_i^d} \left(A_i^k \cap B_i^j\right) \Big/ \left(A_i^k \cup B_i^j\right)}{N_t} \tag{10}$$

where N_t denotes the number of test video, N_i^g and N_i^d denote the number of ground-truth and predicted torsional segments in the *i*th video. A_i^k and B_i^j denote the *k*th detected torsional segment and the *j*th ground-truth torsional segment in the *i*th nystagmus video, respectively.

5.1.3. Implementation

All experiments were performed under a Linux OS on a machine with CPU Intel(R) Xeon(R) E5-2680 v4 @ 2.40GHz, GPU NVIDIA 2080ti. The hyperparameters for training the torsional frame identification model *TBSIN* are shown in Table 4.

5.2. Performance evaluation

5.2.1. Identification of torsional frame

We first evaluate the performance of TBSIN model for torsional frame identification. Here, we use ResNet18 as baseline model but with *x*-optical flow, the *y*-optical flow, and OFTMC as inputs, respectively. As can be seen from Table 5, the proposed bi-stream model TBSIN achieves the best performance in comparison with the other three single-stream based methods. Compared with the *x*-optical flow, it is also clear that *y*-optical flow is more informative in embodying the torsion characteristics of torsional nystagmus. Since the optical flow guided torsion motion confidence map, i.e., OFTMC, has encoded much of information from both *x*-optical flow and *y*-optical flow, it performs better not surprisingly than each of individual *x*-optical flow and *y*-optical flow.

As we know, the activation function in deep learning builds the bridge between the network output layer and the next input layer. Since the optical flow input to the network has both positive and negative values, which is different from the image data that always takes positive

Table 6
Performance comparisons of using different activation functions.

Activation function	*Precision*	*Recall*	*Accuracy*	*F*1
Relu	52.97	65.53	72.39	58.59
Leaky_Relu [44]	58.78	61.78	69.94	60.24
Tanh	62.22	57.14	71.41	59.57
Elu	57.34	68.77	69.62	62.53

Table 7
Performance comparison with and without using Dropout.

	Precision	*Recall*	*Accuracy*	*F*1
w/o Dropout	57.34	68.77	69.62	62.53
w/Dropout	61.60	70.22	72.88	65.62

Table 8
Evaluation on torsional segment detection of nystagmus video.

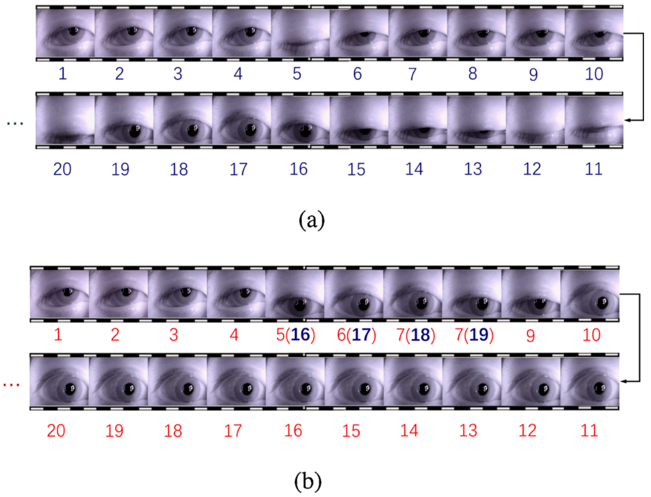
N_w/T	N_b	*Precision*	*Recall*	*Accuracy*	*F*1	IoU_{avg}
4/6	9	82.66	73.48	84.50	77.80	61.32
	11	84.08	73.51	85.06	78.44	60.48
	13	80.99	76.32	84.61	78.59	62.65
	15	83.23	76.22	85.51	79.57	64.72
5/7	9	81.27	76.98	84.93	79.07	64.64
	11	82.69	77.10	85.56	79.79	63.86
	13	79.96	79.77	85.12	79.86	65.75
	15	81.88	78.92	85.73	81.00	67.45
6/9	9	85.05	72.95	85.26	78.54	61.51
	11	85.34	72.27	85.16	78.26	61.26
	13	83.98	76.01	85.76	79.80	63.72
	15	85.14	72.35	85.09	78.23	62.74
7/10	9	83.99	76.04	85.78	79.82	63.38
	11	83.55	76.58	85.77	79.92	64.04
	13	83.76	78.06	86.28	80.80	64.47
	15	83.12	75.18	85.15	78.95	63.39

value, it is necessary to choose an appropriate activation function instead of Relu in Resnet18 when training TBSIN model. Table 6 shows the performance comparisons of using different activation functions. On the whole, Elu performs best compared with the others. This is due to the fact that the average output value of the Elu is close to zero, and it is also more robust to noise than other activation functions.

When training the TBSIN model, we have inserted a Dropout layer in each Basicblock to avoid overfitting during training. As can be seen from Table 7, using the dropout layer in each basicblock does help improve performance. For instance, the F1 score can be increased by 3.09%, while the accuracy is with an improvement by 3.26%.

5.2.2. Localization of Torsional segment

Based on the identification of torsion frames, the complete torsion segments composed of continuous torsional frames can be obtained through frame-level label correction and block-level boundary determination, so as to realize the localization of torsional segments in torsional nystagmus video. Since there are three key parameters N_w, T, and N_b in frame-level label correction and block-level boundary determination, we report in Table 8 the performance variations of detecting the torsional segment by varying these parameters. When $N_w = 5$, $T = 7$, and $N_b = 15$, the optimal IoU for evaluating the performance of torsional segment detection can reach 67.45%. Meanwhile, it should also be noted that the *F*1-Score and Accuracy receive significant increments by 15.38% from 65.62% to 81.00% and 12.85% from 72.88% to 85.73%.



(a)

(b)

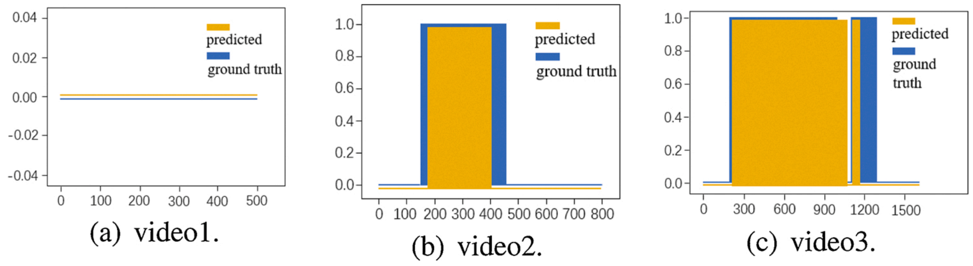
Fig. 15. Result of eye movement video condensation based on CNN. (a) Before condensation; (b) after condensation.

In addition, Fig. 14 also shows the examples of the detected torsion segments in three torsional nystagmus videos, which contain no torsional segment (Fig. 14(a)), 1 torsional segment (Fig. 14(b)), and 2 torsional segments (Fig. 14(c)), respectively. Here, we use the blue area to denote the groundtruth of torsional segment and yellow area the detected torsional segments. As we can observe, there is a high consistency between the groundtruth and the detected torsional segments.

6. Discussions

6.1. Eye movement video condensation and calibration

The aim of video concentration is to eliminate the invalid frames caused by eyelid occlusion, eyebrow interference, blinking, etc., thus facilitating clinicians to make rapid and accurate judgments of the torsion characteristics of the nystagmus. On the other hand, it will also lay the foundation for subsequent pupil calibration. Fig. 15 demosntrates the results before and after condensation of a sequence of eye movement frames. It can be seen that, the invalid frames (frame 5 to 15 and the 20th frame in Fig. 15(a)) cannot provide necessary information to determine the torsion motion of the pupil due to the occlusion of the pupil by the eyelid. Through eye movement video condensation based on convolutional neural network, these invalid frames are effectively eliminated as shown in Fig. 15(b). In practice, we've got a recall of 98%



(a) video1.

(b) video2.

(c) video3.

Fig. 14. Examples of detected torsional segments in 3 nystagmus videos.

Table 9
Performance comparisons of pupil calibration.

Methods	With				Error	Deg. of deviation (%)	Coverage (%)
	CHT	Template matching	Template updating	Morphological pre-processing	(pixel)		
CHT	✓	×	×	×	29.97	0.37	0.47
CHT-TM-1	✓	✓	×	×	14.10	0.17	0.70
CHT-TM-2	✓	✓	✓	×	14.24	0.18	0.71
CHT-TM	✓	✓	✓	✓	7.47	0.09	0.98

Table 10
Video description of two cases of torsional nystagmus.

ID	Originally captured video	Condensed video	Calibrated video	Ground truth of torsional Nystagmus	Detected segments of torsional Nystagmus	IoU (%)
0001-R-PC-BPPV	861 frames	784 frames	555 frames	[54th–273rd] [335th–380th]	[0th–33rd] [55th–303rd] [322nd–381st]	82.13
0002-L-PC-BPPV	949 frames	853 frames	825 frames	[148th-468th]	[184th-393rd]	65.42

on the test dataset.

Pupil calibration is a prerequisite for obtaining optical flow field of nystagmus video. To evaluate the performance of pupil calibration, we define two measures, namely degree of deviation and coverage. Here, the degree of deviation refers to the ratio of the distance between the center of reference pupil and the center of the detected pupil to the radius of reference pupil, and the coverage refers to the ratio of the intersection of the areas of the detected pupil and the reference pupil to the union of them.

We randomly select 10 segments of nystagmus videos, and manually mark the center of the pupil and the corresponding fitting circle in each frame. The experimental results are shown in Table 9, from which we can find that the proposed CHT-TM model achieves an excellent calibrating performance compared with the other methods. CHT denotes that only using traditional Circular Hough Transform algorithm to locate the pupil center; CHT-TM-1 and CHT-TM-2 indicate combining CHT and template matching but without or with template updating to be considered.

6.2. Case study

In Table 10, we finally present the descriptions for two cases of captured torsional nystagmus videos, which are left torsional PC-BPPV and right torsional PC-BPPV, respectably. Taking 0001-R-PC-BPPV as an example, the length of originally captured torsional nystagmus video is reduced from 841 frames in total to 555 frames, which means nearly 34% of non-informative frames are effectively eliminated. Meanwhile, we can also notice that the first located segment of torsional nystagmus (i.e., [0th–33rd] frames) is indeed a false detection. It is mainly due to the unstable eye movement of the patient during the initial video capture phase. Additionally, the proposed model achieves 82.13% performance on IOU for this case. For the convenience of reviewers' reference, these two cases are uploaded to the submission platform as supplementary materials.

7. Conclusions

Torsional nystagmus is the most diagnostic signs associated with benign paroxysmal positional vertigo (BPPV). Thus, it is of great clinical significance to realize the automatic detection of torsional nystagmus, which cannot only relieve doctors' working burden effectively, improve their diagnostic efficiency, but also provide support for the final auxiliary diagnostic system. We mainly in this paper focus on developing a deep learning based framework for torsional nystagmus detection. The proposed pipeline framework is composed of eye movement video condensation and calibration, establishing the optical flow guided torsion motion confidence (OFTMC) map, and deep convolutional network based torsional nystagmus detection. In particular, a novel torsion-aware bi-stream identification network (TBSIN) model was proposed to perform frame-level recognition of torsion motion, on the basis of which the torsional segments can be automatically localized. The collected dataset about pc-BPPV in this work not only validates the effectiveness of our proposed approach, but also can serve as a good benchmark for research community.

CRediT author statement

Wanlu Zhang: Investigation, Methodology, Visualization, Writing – Original Draft
Haiyan Wu: Conceptualization, Data Curation, Project administration, Writing – Review & Editing
Yang Liu: Formal analysis, Validation
Shuai Zheng: Software, Validation
Zhizhe Liu: Software
Youru Li: Software
Yao Zhao: Project administration
Zhenfeng Zhu: Supervision, Formal analysis, Writing – Review & Editing

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

This work was supported in part by Science and Technology Innovation 2030 – New Generation Artificial Intelligence Major Project under Grant 2018AAA0102101, in part by the National Natural Science Foundation of China under Grant No. 61976018.

References

[1] L.T. Roland, D. Kallogjeri, B.C. Sinks, et al., Utility of an abbreviated dizziness questionnaire to differentiate between causes of vertigo and guide appropriate referral, Otol. Neurotol. 36 (10) (2015) 1687–1694.
[2] M. Bethesda, Nih:a report of the task force on the national strategic research plan 74.
[3] P.D. Sloane, Dizziness in primary care. results from the national ambulatory medical care survey, J. Fam. Pract. 29 (1) (1989) 33–38.
[4] M. Guerraz, A.M. Bronstein, Ocular versus extraocular control of posture and equilibrium, Neurophysiol. Clin./Clin. Neurophysiol. 38 (6) (2009) 391–398.

[5] A. Aydemir, A. Uneri, Detection and analysis of quick phase eye movements in nystagmus (vng), 2006 IEEE 14th Signal Processing and Communications Applications (2006).

[6] A.B. Slama, A.N. Machraoui, M. Sayadi, Pupil tracking using active contour model for video nystagmography applications, International Image Processing, Applications and Systems Conference (2014).

[7] S.D.Z. Eggers, A. Bisdorff, M.V. Brevern, Classification of vestibular signs and examination techniques: nystagmus and nystagmus-like movements: consensus document of the committee for the international classification of vestibular disorders of the bárány society, J. Vestib. Res. 29 (2,3) (2019) 1–31.

[8] M. Von Brevern, P. Bertholon, T. Brandt, et al., Benign paroxysmal positional vertigo: diagnostic criteria, J. Vestib. Res. Equilib. Orientat. 25 (3,4) (2015) 105–117.

[9] L. Fang, B. Peng, W. Liu, et al., New feature of nystagmus and its application in benign pistional paroxysmal vertigo, IEEE International Conference on Awareness Science & Technology (2015).

[10] L.S. Parnes, S.K. Agrawal, J. Atlas, Diagnosis and management of benign paroxysmal positional vertigo (bppv), Can. Med. Assoc. J. 169 (7) (2003) 681.

[11] M.L, V.W, M.v, Classification of benign paroxysmal positioning vertigo types from dizziness handicap inventory using machine learning techniques, 2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), vol. 3 (2018).

[12] D.K. Kumar, A. Banerjee, S. Swaminathan, et al., Mems modeling of the posterior semicircular canal for treating benign paroxysmal positional vertigo, The 8th Annual IEEE International Conference on Nano/Micro Engineered and Molecular Systems (2013).

[13] N. Bhattacharyya, S.P. Gubbels, S.R. Schwartz, et al., Clinical practice guideline: benign paroxysmal positional vertigo (update), Otolaryngol.-Head Neck Surg. 156 (3_suppl) (2017) S1–S47.

[14] W. Cong, J. Zhang, L. Niu, et al., Dovenet: deep image harmonization via domain verification, 2020 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2020).

[15] J.U. Kim, Y. Man Ro, Attentive layer separation for object classification and object localization in object detection, 2019 IEEE International Conference on Image Processing (ICIP) (2019).

[16] S. Mane, S. Mangale, Moving object detection and tracking using convolutional neural networks, 2018 Second International Conference on Intelligent Computing and Control Systems (2018).

[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).

[18] M. Sundermeyer, H. Ney, R. Schluter, From feedforward to recurrent lstm neural networks for language modeling, IEEE/ACM Trans. Audio Speech Lang. Process. 23 (3) (2015) 517–529.

[19] T. Young, D. Hazarika, S. Poria, et al., Recent trends in deep learning based natural language processing [review article], IEEE Comput. Intell. Mag. 13 (3) (2018) 55–75.

[20] A. Karpathy, F. Li, Deep visual-semantic alignments for generating image descriptions, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 664–676.

[21] M. Anthimopoulos, S. Christodoulidis, L. Ebner, et al., Lung pattern classification for interstitial lung diseases using a deep convolutional neural network, IEEE Trans. Med. Imaging 35 (5) (2016) 1207–1216.

[22] A.A.A. Setio, F. Ciompi, G. Litjens, et al., Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks, IEEE Trans. Med. Imaging 35 (5) (2016) 1160–1169.

[23] Z. Wu, S. Zhao, Y. Peng, et al., Studies on different cnn algorithms for face skin disease classification based on clinical images, IEEE Access 7 (2019) 66505–66511.

[24] J. Rathod, V. Waghmode, A. Sodha, et al., Diagnosis of skin diseases using convolutional neural networks, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (2018).

[25] D.S. Ting, Y.L. Cheung, G. Lim, et al., Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes, J. Am. Med. Assoc. 318 (22) (2017) 2402–2410.

[26] M.E. Sertkaya, B. Ergen, M. Togacar, Diagnosis of eye retinal diseases based on convolutional neural networks using optical coherence images, 2019 23rd International Conference Electronics (2019).

[27] P.N.H. Tra, N.T. Hai, T.T. Mai, Image segmentation for detection of benign and malignant tumors, 2016 International Conference on Biomedical Engineering (BME-HUST) (2016).

[28] R. Lavanyadevi, M. Machakowsalya, J. Nivethitha, et al., Brain tumor classification and segmentation in mri images using pnn, 2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (2017).

[29] Y. Lecun, B. Boser, J. Denker, et al., Backpropagation applied to handwritten zip code recognition, Neural Comput. 1 (4) (1989) 541–551.

[30] J. Ker, L. Wang, e.a.J. Rao, Deep learning applications in medical image analysis, IEEE Access 6 (2018) 9375–9389.

[31] S.V.M. Kumar, R. Nishanth, N. Sani, et al., Specular reflection removal using morphological filtering for accurate iris recognition, 2019 International Conference on Smart Structures and Systems (ICSSS) (2019).

[32] R.H. Nugroho, M. Nasrun, C. Setianingsih, Lie detector with pupil dilation and eye blinks using hough transform and frame difference method with fuzzy logic, 2017 International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC) (2017).

[33] P. Bonteanu, R.G. Bozomitu, A. Cracan, et al., A new pupil detection algorithm based on circular hough transform approaches, 2019 IEEE 25th International Symposium for Design and Technology in Electronic Packaging (SIITME) (2019).

[34] H. Zhang, Multiple moving objects detection and tracking based on optical flow in polar-log images, 2010 International Conference on Machine Learning and Cybernetics, vol. 3 (2010).

[35] Y. Chen, Q. Wu, Moving vehicle detection based on optical flow estimation of edge, 2015 11th International Conference on Natural Computation (ICNC) (2015).

[36] Z. Wang, X. Yang, Moving target detection and tracking based on pyramid lucas-kanade optical flow, 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC) (2018).

[37] S. Wang, X. Shen, J. Liu, Dense optical flow variation based 3d face reconstruction from monocular video, 2018 25th IEEE International Conference on Image Processing (ICIP) (2018).

[38] A. Lowhur, M.C. Chuah, Dense optical flow based emotion recognition classifier, 2015 IEEE 12th International Conference on Mobile Ad Hoc and Sensor Systems (2015).

[39] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, in: Scandinavian Conference on Image Analysis, Springer, 2003, pp. 363–370.

[40] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.

[41] M. Siam, M. Gamal, M. Abdel-Razek, et al., Rtseg: Real-time semantic segmentation comparative study, 2018 25th IEEE International Conference on Image Processing (ICIP) (2018).

[42] T.N. Dahl, G.E. Sainath, Hinton, Improving deep neural networks for lvcsr using rectified linear units and dropout, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (2013).

[43] A. Ashiquzzaman, A. Tushar, S. Dutta, et al., An efficient method for improving classification accuracy of handwritten bangla compound characters using dcnn with dropout and elu, 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (2017).

[44] Y.D. Zhang, X.X. Hou, Y. Chen, et al., Voxelwise detection of cerebral microbleed in cadasil patients by leaky rectified linear unit and early stopping, Multimed. Tools Appl. 77 (17) (2017) 21825–21845.