Recurrent Fully Convolutional Networks Based on Optical Flow for Video Eyes Fixation Prediction

Jiu-chen SHI¹ and Dong ZHANG²,*

¹School of Computer Software, Dalian University of Technology, Dalian 116620, China
²Institute of Chinese Medicine, Tianjin University of Traditional Chinese Medicine, Tianjin 300193, China
*Corresponding author

Keywords: Recurrent fully convolutional neural network, Optical flow, Eye fixation, Prediction.

Abstract. Although the research on eye fixation prediction has been activated in recent years, most of the methods are on images and cannot be directly applied to videos. In this paper, a recurrent fully convolutional neural network (RFCN) based on the optical flow framework was proposed for predicting the eye fixation. Seven frames and their optical flow feature were used as the input. We used the RFCN without deconvolution layers to extract the frames and their optical flow of the video to obtain the convolution features. Adjacent convolution features were merged as the next input. A Gaussian blob emphasized the center tendency of human eyes fixation when calculating the convolutional features. Finally, VAGBA and CRCNS public video databases were used to evaluate the model. The experimental results show that the prediction performance is better than the contrastive algorithms.

Introduction

Human eyes visual fixation prediction is an important issue of video content analysis [1]. It has a significant use in semantic analysis, target tracking, object segmentation and other aspects [2]. Therefore, predicting the visual fixation area by using an automatic and accurate algorithm in the video can effectively assist people in video analysis [3].

In recent years, many work had been done on studying human eye fixation attention from the static images, but it has a lot of limit when predicting in the video [4]. It is because when watching the video, the observer’s attention is highly dynamic [5]. Thus, existing image prediction models cannot be directly applied to videos. For videos, the most common issue was feature extraction. Low and high level feature had been extracted such as color, objects and actions. However, these existing methods had some shortages. The main reason is that the handcrafted features are complex and difficult to decide which is useful. And the motion and temporal factor were not considered between the adjacent frames [6]. Recently, deep learning models have shown good results on visual process. Fully convolutional neural network (FCN) was currently widely used in image segmentation, target detection and so on. It has achieved good results in the field of saliency marking and visual fixation detection [7]. The fully convolutional neural network converts the fully connecting layer in a traditional convolutional neural network into multiple convolutional layers and obtains the same output size as the input through deconvolution and up-sampling operation. Recurrent fully convolutional network (RFCN) was an iterative algorithm based on FCN. In each time slot, an input RGB image and a priori significant mapping are provided by the RFCN to obtain a significant mapping of the prediction. This mapping is used again as a significant priori map in the next time slot.

Different from the traditional methods, this paper focuses on the eye fixation prediction with RFCN and optical flow. In the framework, a fully convolutional neural network without deconvolution layers was used to the single frame and the optical flow map, about seven frames were extracted as the input. The moving objects can attract the viewer’s more attention and the central region of the frame also attracts more attention of human’s eyes. The convolution feature of the previous frame was
calculated with the current frame’s convolution feature to obtain the new fusion feature, and we called it recurrent fully convolutional neural network. We summarize the key aspects of our hybrid network:

1) A recurrent fully convolutional neural network with several input frames were used in the video eye fixation prediction.

2) Optical flow was used to capture the moving objects feature of the frames, the Gaussian blob emphasize the center tendency of human eyes fixation.

**The Proposed Model**

The framework of the recurrent fully convolutional neural network based on the optical flow model was shown in Fig. 1. Firstly, a few of video frames was taken as the input to the proposed model. Through the RFCN processing, we can get the temporal and spatial feature based on RFCN of the frames and its optical flow. Eye tracking data was calculated to obtain the fixation map as the groundtruth. Finally, for the frames in the testing videos, the eye fixation prediction value can be calculated which shown as the saliency map. The training and testing process was done on two public eye fixation video datasets.

![Figure 1. The framework of the proposed method.](image)

In recent years, deep learning models had been shown a good performance in many fields, such as tracking, recognition and so on. It can learn hierarchy of features from low-level features to high-level features.

In this paper, the fully convolutional neural network without deconvolution layers was used in every input frame. The architecture of the single frame FCN can be seen in Fig. 2. Finally, we can obtain the convolution process feature of the single frame.

![Figure 2. The architecture for the single frame image about the fully convolution network without the deconvolution layers.](image)

Statistically, the central region of the frame in the video can attract more attention of human’s eyes. This phenomenon was called the central prejudice in eye gaze and conducted extensive research in
neuroscience and psychology. Thus a Gaussian blob (Top right in Fig. 3) was calculated with the first convolution layer feature to emphasize the characteristics of human eyes fixation.

The recurrent network was shown in Fig. 3. As can be seen, the current frame as its prior six frames in the video were taken as the input to the proposed model. The optical flow was calculated on continuous frames about all the input frames, which can obtain the temporal moving features that catching the eye attention in video. Fully convolutional neural network without the deconvolution layers was used in the prior six frames to get the convolution feature. And then, the convolution feature of the previous frame was calculated with the current frame’s convolution feature to obtain the new fusion feature. Through the continuous iteration, the recurrent convolutional neural network was formed. What’s more, a Gaussian blob was used in the initial convolution feature. Finally, deconvolution (deconvolution layer 1 (8×8×1 stride 4)) was done to obtain the eye fixation prediction map.

Results

Video Databases

In this paper, we used the public video database VAGBA [8] and CRCNS [9] to evaluate the accuracy of the proposed model. The VAGBA video database contains 50 video clips which are about 10 seconds. The CRCNS video database contains 50 video clips with duration of 20-60 seconds.

In order to evaluate the accuracy of this method, we also need to obtain the ground truth map of human eyes in the video database. Each video database used in this paper independently records the position coordinates of the human eyes of different observers watching each frame of video through professional instruments. The VAGBA records the eye's focus of 14 observers, and the human eyes fixation changes in the CRCNS video database is marked by 8 observers. The ground truth that marked the probability of a pixel to be fixated of the video based the 3D Gaussian model was calculated [10].

Experiment Setup

In this paper, the two databases, VAGBA and CRCNS, were used to randomly select 80% of the clips in each video database as the training set for the training of the model. The remaining 20% was used
as a testing dataset for cross-validation to test the accuracy of this method. For the FCN model, we trained over 200K iterations and set the initial learning rate to $10^{-7}$, the momentum to 0.9 and the weight decay to 0.0005.

**Results**

In order to evaluate the proposed model, we use Earth Mover’s Distance, Area Under Receiver Operating Characteristic to evaluate the performance of the model. Many commonly used approached about the video eye fixation prediction were used as the compared experiments to evaluate the performance of the proposed model. They were GBVS[11], SR [12], Rahtu [13], MLSN[14] and STAD[15]. The results can be seen in Table 1.

In the EMD results, the proposed method achieved the smallest value which mean that the prediction saliency map was closer with the groundtruth. In the AUC, the proposed method performs better than others in most cases. Therefore, the proposed method obtains a better performance than the compared methods.

<table>
<thead>
<tr>
<th></th>
<th>Our Method</th>
<th>GBVS</th>
<th>SR</th>
<th>Rahtu</th>
<th>MLSN</th>
<th>STAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMD (VAGBA)</td>
<td>0.4247</td>
<td>0.4505</td>
<td>0.4396</td>
<td>0.4468</td>
<td>0.4337</td>
<td>0.4382</td>
</tr>
<tr>
<td>EMD (CRCNS)</td>
<td>0.6623</td>
<td>0.7125</td>
<td>0.6932</td>
<td>0.6845</td>
<td>0.6695</td>
<td>0.7073</td>
</tr>
<tr>
<td>AUC(VAGBA)</td>
<td>0.8303</td>
<td>0.8121</td>
<td>0.8153</td>
<td>0.8085</td>
<td>0.8233</td>
<td>0.8187</td>
</tr>
<tr>
<td>AUC(CRCNS)</td>
<td>0.6926</td>
<td>0.6637</td>
<td>0.6741</td>
<td>0.6773</td>
<td>0.6807</td>
<td>0.6771</td>
</tr>
</tbody>
</table>

In order to clearly shown the performance of our method, we randomly calculated some video frames with the ground truth and the prediction eye fixation map. As shown in Fig. 4, all the prediction results were illustrated and the moving objects were attracting more human’s attention. Different color represents different attention of the human eye in all the prediction saliency map. As can be seen, the red color represents the most attention, the yellow and blue followed. In most cases, the results of the proposed method matched the groundtruth best. For example, the ‘Garden07’ in the VAGBA video dataset, the result of the proposed method was more accurate in the region of saliency and the location was more concentrated. In other methods, the area of the prediction is larger, more widely distributed, and the result is less accurate.

**Conclusion**

In this paper, we proposed a RFCN model based on the optical flow to predict the eye fixation map. Seven frames and their optical flow feature were used as the input. A RFCN model with a Gaussian blob were used to calculate the saliency map. Experiment results showed that the proposed method
was effective than the comparative method in the two public video datasets. In the future, we will continue to improve the RFCN model, and consider more human visual characteristics. Moreover, pre-processing and more temporal relation analysis in video will be used to solve the problem.

Acknowledgement
This research was financially supported by the Tianjin Municipal Education Commission research project (No. 2017KJ151).

References
[10] J. Han, L. Sun, X. Hu, J. Han, L. Shao, Spatial and temporal visual attention prediction in videos using eye movement data, Neurocomputing. 145 (2014) 140-153.