Deep learning on natural viewing behaviors to differentiate children with fetal alcohol spectrum disorder

Po-He Tseng[1], Angelina Paolozza[2], Douglas P. Munoz[2], James N. Reynolds[2], and Laurent Itti[1]

[1] Department of Computer Science, University of Southern California, USA
[2] Centre for Neuroscience Studies, Queen's University, Canada

Abstract. Computational models of visual attention have attracted strong interest by accurately predicting how humans deploy attention. However, little research has utilized these models to detect clinical populations whose attention control has been affected by neurological disorders. We designed a framework to decypher disorders from the joint analysis of video and patients' natural eye movement behaviors (watch television for 5 minutes). We employ convolutional deep neural networks to extract visual features in real-time at the point of gaze, followed by SVM and Adaboost to classify typically developing children vs. children with fetal alcohol spectrum disorder (FASD), who exhibit impaired attentional control. The classifier achieved 74.1% accuracy (ROC: 0.82). Our results demonstrate that there is substantial information about attentional control in even very short recordings of natural viewing behavior. Our new method could lead to high-throughput, low-cost screening tools for identifying individuals with deficits in attentional control.

Keywords: deep learning, sparse representation, eye movement, natural scenes, clinical populations

1 Introduction

Attention enables us to interact with complex environments by selecting relevant information to be processed in the brain. Eye movements have served as a probe for visual attention in a wide range of conditions, because both are strongly coupled in our everyday life [8]. A large brain network is recruited by attention, which makes attention control susceptible to neurological disorders [1]. Thus, psychologists have designed eye-tracking experiments with simple artificial stimuli to characterize the influence of these disorders on eye movement metrics and to assist diagnosis [8].

In the past decades, computational models of visual attention have shown significant ability to predict human eye movement while people view complex natural scenes and perform complex tasks. The success of these attention models, combined with eye tracking data, allows us to assess how neurological disorders impact attention under a more natural scenario, compared to traditional psychophysical experiments with artificial stimuli. Previous research has utilized attention models to assess autism [2] and agnosia [11, 3], however, without much success (null or inconsistent results). This is possibly due to the complexity of natural scenes and natural viewing eye traces, especially when the differences between patients and controls are subtle.

In this study, we propose a novel method that allows us to use only 5 minutes of natural viewing eye traces to identify children with FASD. FASD is caused by prenatal exposure to alcohol, which impacts the brain globally and impairs attentional control. Our new method computes "attention traces" by collecting the predictions of attention models (saliency values) along recorded human eye movement traces. To represent these complex attention traces, we utilize a convolutional deep neural network [9] to learn their sparse representation in an unsupervised manner, and to discover the features that best capture variations in attention traces; we then differentiate children with FASD from typically developing children by classifying these features over time.

2 Methods

2.1 Paradigm overview

Participants' eye movements were recorded while they watched five 1-minute videos, and they were instructed to "watch and enjoy the clips". These clips were composed of content-unrelated clip snippets of 2-4 seconds each (a few were 7-8 seconds). This design attempted to reset the observers' attention status with each new snippet (Fig. 1A). For each video frame, Itti's visual attention model was used to compute a saliency map that predicts visually conspicuous locations from low-level visual attributes of the video frame. The normalized map values were extracted along the eye trace (attention trace). To extract features from the attention traces, a convolutional deep neural network was utilized to discover 256 bases from the attention traces of control young adults. Next, we applied these bases on the child data and obtained a sparse representation of their attention traces. This representation was input to a L1-regularized logistic SVM classifier after initial filter feature selection. We built a weak classifier from each clip snippet independently, and used Adaboost to construct a strong classifier for the final prediction.

2.2 Computational model of visual attention

To evaluate an observer's bottom-up attention, we used the popular Itti *et al.* saliency model [7, 6, 10] to identify visually salient locations that may capture attention in natural videos. Itti's model was biologically inspired to mimic early visual processing in humans.

To estimate visual salience in a scene, the saliency model (Fig. 1B) first applied a linear filter for each of several visual attributes (e.g., color, intensity)

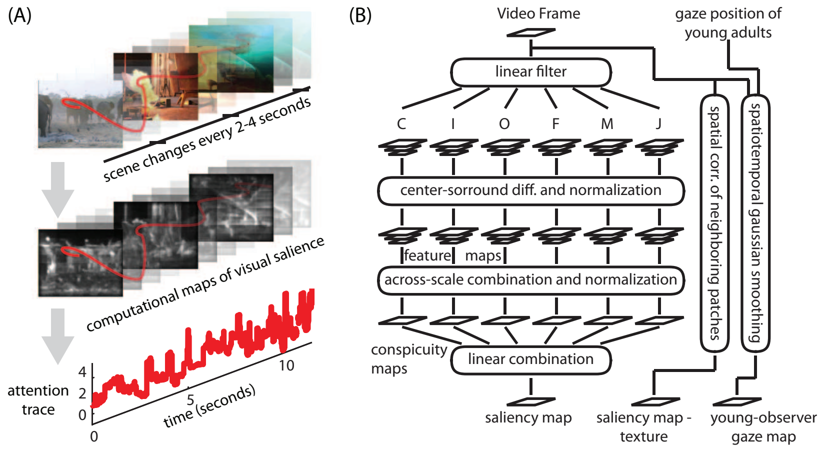
(A)
scene changes every 2-4 seconds
computational maps of visual salience
attention trace
4
2
0
0
5
10
time (seconds)
(B)
Video Frame
gaze position of young adults
linear filter
C
I
O
F
M
J
center-sorround diff. and normalization
feature maps
across-scale combination and normalization
spatial corr. of neighboring patches
spatiotemporal gaussian smoothing
conspicuity maps
linear combination
saliency map
saliency map - texture
young-observer gaze map

Fig. 1. Evaluating attention deployment. (A) To extract an attention trace, observers' eye movements were recorded (red curve) during free viewing of videos of natural scenes. For each corresponding video frame, ten different types of saliency maps were generated. Brighter intensity in a map indicates locations of the video frame with stronger feature contrast. The normalized map values at every gaze location were extracted to yield the attention trace. (B) Architecture of Itti's computational model of visual attention, extended here to contain many more feature channels than the original implementation. C, color; I, intensity; O, orientation; F, flicker; M, motion; J, line junction.

on a video frame, at several scales to generate multi-scale (i.e., fine to coarse) filtered maps. Fine filtered maps were then subtracted from coarse filtered maps to simulate center-surround operations in human vision and to produce feature maps. Next, these multi-scale feature maps were normalized and combined together to generate conspicuity maps in a manner that favors feature maps with sparse peak responses. Finally, conspicuity maps of different features were linearly summed together to form a saliency map, if multiple visual attributes were evaluated together. Equations of these linear filters were described in [7, 6, 10].

Ten types of maps used in this study provided information on attention control. Nine saliency maps of different visual attributes estimate bottom-up, stimulus-driven attention, and one top-down map, generated by instantaneous gaze positions of 19 normative young adults, provides information on voluntary, top-down attention control in addition to bottom-up attention. Out of the 9 saliency maps, 7 were individual visual attributes (color, intensity, orientation, flicker, motion, line junction, and texture [10, 12]), and 2 were combination of multiple visual attributes (CIOFM and CIOFMJ; Fig. 1B for abbreviations). For each map, we quantified the agreement between eye movement traces and the map, and such agreement over time yielded attention traces (see Supplementary

Methods[1]). With the ten maps, ten attention traces were obtained. While eye movements were recorded at 500Hz, attention traces were resampled to 125Hz to reduce the computational load. Next, the 10 attention traces were whitened by principal component analysis (PCA), and the first 7 principal components were kept to maintain 97.5% of the information.

2.3 Convolutional Deep Neural Network

We chose a two-layer convolutional deep neural network (CDNN) [9] to discover the sparse representation of the attention traces. The sparse representation had several advantages in representing input signals, and the one which benefited classification the most was denoising, because signals and noise were likely to be separated by different components of the representation. Each layer of the CDNN was a topographic independent component analysis (TICA) network (Fig. 2) [5]. TICA discovers components (bases) from unlabelled data in a way similar to independent component analysis (ICA). However, TICA features a loosened independence constraint between neighboring components so that similar components were next to each other, which gave rise to a more stable set of independent components [5].

Our TICA network was composed of three components: input, simple, and pooling nodes (Fig. 2A). Given an input pattern $x^{(t)}$, each pooling node was activated as $p_i(x^{(t)}; W, V) = \sqrt{\sum_{k=1}^{m} V_{ik} (\sum_{j=1}^{n} W_{kj} x_j^{(t)})^2}$, where n was the local receptive field size (size of the input) and m was the number of simple units. The weights $V \in \mathbb{R}^{m \times m}$ between the simple and pooling nodes were fixed ($V_{ij} = 0 \, or \, 1$) to represent to topographical structure. The weights $W \in \mathbb{R}^{m \times n}$ between the input and simple nodes were the sparse bases to be learned by TICA. The TICA learned the weights W by solving

$$\arg\min_{W} \sum_{t=1}^{T} \sum_{i=1}^{m} p_i(x^{(t)}; W, V), \quad subject\ to \quad WW^T = I \tag{1}$$

where input patterns $\{x^{(t)}\}_{t=1}^{T}$ were whitened attention traces. This objective function minimized the sum of activities from all pooling nodes to reduce redundancies between them, and the constraint $WW^T = I$ ensured sparseness between bases. This objective function corresponded to classic ICA if each pooling node p_i was connected to exactly one simple node h_i.

The sparse representation of children's attention traces was simply obtained by propagating their whitened traces through the CDNN pretrained from the young adult controls. The children's attention traces were whitened by the same principal components that young adult controls used. These whitened traces then propagated through the CDNN composed of two TICA networks. For each TICA network, the output of each basis along the whole attention trace was called a "trace-map", and all the trace-maps together should cover the entire attention

[1] available at http://ilab.usc.edu/publications/doc/tseng_etal13idealsupp.pdf

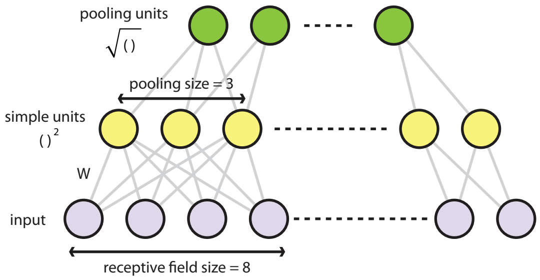

Fig. 2. One TICA layer (itself consisting of 3 layers) of the convolutional deep neural network, where the simple units are the squared weighted sum of the receptive field, and the pooling units compute the square root of the sum of adjacent simple units.

trace. We concatenated all the trace-maps into one large vector that sparsely represented the saliency traces and would be the features used for classification (see Supplementary Methods for detail).

2.4 Classification Procedures

A weak support vector machine (SVM) classifier was built per clip snippet, and a strong classifier was constructed by Adaboost [4], from the weak classifiers that performed better than the chance level ($\frac{max(n_i)}{\sum n_i}$, n_i: number of participants of group i). For each weak classifier, the feature dimension (all the trace-maps) was much higher than the number of training samples, e.g., a clip snippet of 3.129 seconds yields 17,408 features. To reduce feature dimensions, a combination of filter and wrapper methods was used in selecting features that discriminated patients from controls in the training data set. For the filter method, we used a two-tail t-test to filter features whose p-value ≤ 0.05 after Bonferroni correction within each trace-map. Once the filter method selected a set of features, an L1-regularized logistic support vector machine (SVM) was used to perform another feature selection and classification simultaneously, and resulted in our weak classifier, which used 33 out of the 17,408 features of this 3.129 second clip snippet. A 5-fold cross validation was used on the training dataset to determine the cost value (10, 1, 0.1, 0.0001) of the SVM.

Within the training dataset, we performed a leave-one-out cross validation (LOOCV) to estimate the generalizability of each weak classifier to different training data, and to estimate the consistency of predictions on test data. The LOOCV left one participant per group out. A weak classifier would be recruited in the Adaboost strong classifier if it met 2 criteria: (1) the weak classifier performed above chance level in the training dataset (generalizability), and (2) the testing data were predicted consistently during the LOOCV. To estimate the consistency, we took the median of the predicted class probability throughout the LOOCV. If the median was higher than 0.7 or below 0.3, then the weak classifier was considered consistent. Moreover, the median predicted class label was the prediction that served as the input of the strong Adaboost classifier.

LOOCV was also used to test the performance of the strong classifier. We first left one participant out. With the remaining participants, we performed the filter feature selection, trained the L1-regularized SVM, selected the weak classifiers based on their generalizability, and trained the adaboost classifier. Finally, the learned classifier was used on the participant who was left out at the beginning. The LOOCV was repeated 10 times with different fold structures. Note that we did not balance the number of participants in each group because our data were only slightly unbalanced (53 controls vs. 47 patients).

3 Experimental Evaluation

3.1 Data

The results reported in this study included 53 control children (11.11±3.31 yr) and 47 children with FASD (12.21±3.32 yr). An additional nineteen young adult controls (20.74±1.33 yr) were recruited to collect data to train the sparse representation of the attention trace by CDNN. The participants watched high-definition (1280×1024 pixels) continuous natural scene videos whose content changed every 2-4 seconds (61 snippets) or 7-8 seconds (9 snippets). Each video (about 1 minute) included 13 to 15 snippets. Participants were instructed to "watch and enjoy the clips", and their right eye was tracked at 500Hz when they watched the videos (see Supplementary Experiment for detail).

3.2 Results

We compared the classification performance of the classifier using either (1) the gaze positions in screen coordinates, (2) the 10 raw attention traces, or (3) the attention traces in sparse representations. All 3 test cases used a t-test filter, L1-regularized SVM, and the adaboost classifier. They only differed in the representation of the eye traces. These classifiers achieved accuracies of 49.3%, 53.4%, and 74.1% respectively (Fig. 3A). Wilcoxon signed-rank tests showed that the classifier using the sparsely represented attention traces outperformed the classifiers using gaze positions and raw attention traces ($p<0.01$). Furthermore, the classifier using raw attention traces performed slightly better than that using gaze positions ($p=0.04$). The confusion matrix of the classifier with sparsely represented attention traces (Fig. 3B) showed precision 73.9%, recall (sensitivity) 69.4%, and specificity 78.3% on average.

To perform receiver operating characteristic (ROC) analysis, decision values of the Adaboost classifier were normalized and converted to probabilities based on Bayes rules. The ROC analysis showed that the classifier with sparse attention traces (area under the curve, AUC=0.82) outperformed the classifiers with gaze positions (AUC=0.49) and with raw attention traces (AUC=0.56) ($p<0.01$). Moreover, the classifier with raw attention traces performed better than the classifier using gaze positions ($p=0.02$).

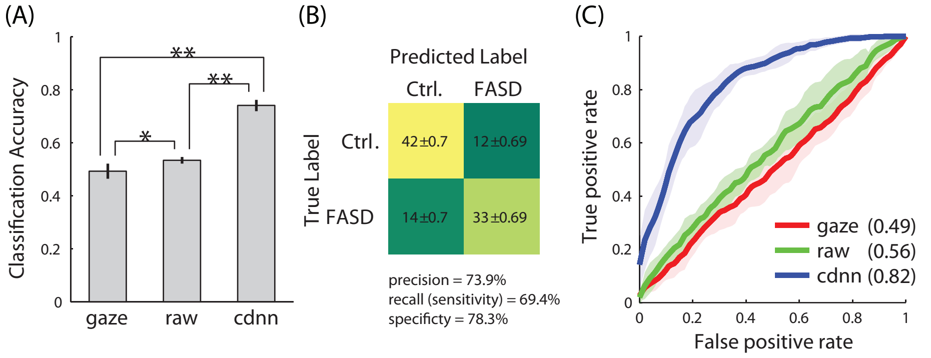

Fig. 3. (A) Classification accuracies of the classifiers using the gaze positions (*gaze*), the raw attention traces (*raw*), and the attention traces in sparse representation learned by convolutional deep neural network (*cdnn*). *, p<0.05; **, p<0.01 (B) Confusion matrix of the classifier using the sparsely represented attention traces. (c) ROC analysis. The numbers inside parentheses are area under the curve for each classifier.

3.3 Comparisons

Diagnosing FASD currently is a complex and time-consuming process. Several studies have attempted to assist diagnosis by building classification rules that are based on objective measures on how FASD influences children's brain structure and behaviors. However, most of such studies used the same dataset to train and test their classifier, and some studies used the same dataset to select features and to train the classifier (double-dipping) (see Supplementary Comparisons). Therefore, their classification performance may be overestimated.

One fair comparison is the study done by Tseng et al. [12]. Similar to our study, they investigated 15 minutes of natural viewing behavior combined with Itti's computational model of visual attention to classify children with FASD from controls. They reported 79.2% classification accuracy by leave-one-out cross validation, which was comparable to our results. However, their work was based on only one type of eye movement, saccades, which occured only ∼1.5 times per second on average for children. The low saccade frequency may limit their method to be further shortened, because small numbers of saccades may not provide reliable statistics to describe attention deployment. In fact, we used their algorithm (provided by the authors) on our 5 minutes of data, and the classification accuracy was 54.1%, which was lower than ours (74.1%).

4 Conclusions and Future Work

This study demonstrates that substantial information about the state of one's attention control exists in one's natural viewing behavior. By combining computational models of visual attention and convolutional deep neural networks, such information can be successfully decoded. With such a short experiment and the low cost of eye trackers, this method shows potential to be used as a screening tool that can be broadly deployed (unlike, e.g., MRI).

This study serves as the first step in decoding attention control status from natural behavior, and several extensions of this work can be explored. For example, video stimuli can be designed and mined to specifically discriminate different populations or disorders. Different representations of the attention traces can be investigated, and each attention trace for each feature can be separately classified to give rise to a multi-dimensional diagnostic profile (biometric signature). Correlating the classifier to standard behavioral tests and functional/structural imaging data can advance our understanding of what the discriminative information is which the classifier learned. The underlying generating processes of attentional control can be modelled to better understand attentional control. In summary, this study opens up a new research area that is exciting to explore.

References

1. Corbetta, M., Patel, G., Shulman, G.L.: The reorienting system of the human brain: from environment to theory of mind. Neuron **58**(3) (May 2008) 306–24
2. Fletcher-Watson, S., Leekam, S.R., Benson, V., Frank, M.C., Findlay, J.M.: Eye-movements reveal attention to social information in autism spectrum disorder. Neuropsychologia **47**(1) (January 2009) 248–57
3. Foulsham, T., Barton, J.J.S., Kingstone, A., Dewhurst, R., Underwood, G.: Modeling eye movements in visual agnosia with a saliency map approach: bottom-up guidance or top-down strategy? Neural networks **24**(6) (August 2011) 665–77
4. Freund, Y., Schapire, R.: A desicion-theoretic generalization of on-line learning and an application to boosting. Computational learning theory **904**(1) (1995) 23–37
5. Hyvärinen, a., Hoyer, P.O., Inki, M.: Topographic independent component analysis. Neural computation **13**(7) (July 2001) 1527–58
6. Itti, L., Dhavale, N., Pighin, F.: Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention. In Bosacchi, B., Fogel, D.B., Bezdek, J.C., eds.: Proc. SPIE 48th Annual International Symposium on Optical Science and Technology. Volume 5200., Bellingham, WA, SPIE Press (August 2003) 64–78
7. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11) (November 1998) 1254–1259
8. Karatekin, C.: Eye tracking studies of normative and atypical development. Developmental Review **27**(3) (September 2007) 283–348
9. Le, Q.Q., Ngiam, J., Chen, Z., hao Chia, D.J., Koh, P.W., Ng, A.Y., Chia, D.: Tiled convolutional neural networks. In Lafferty, J., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A., eds.: Advances in Neural Information Processing Systems 23. (2010) 1279–1287
10. Li, Z., Itti, L.: Saliency and Gist Features for Target Detection in Satellite Images. IEEE transactions on image processing (December 2010)
11. Mannan, S.K., Kennard, C., Husain, M.: The role of visual salience in directing eye movements in visual object agnosia. Current biology **19**(6) (March 2009) R247–8
12. Tseng, P.H., Cameron, I.G.M., Pari, G., Reynolds, J.N., Munoz, D.P., Itti, L.: High-throughput classification of clinical populations from natural viewing eye movements. Journal of neurology (August 2012)