Neural Networks 105 (2018) 52–64

Contents lists available at ScienceDirect

Neural Networks

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

# Characterization of electroencephalography signals for estimating saliency features in videos

Zhen Liang [a,*], Yasuyuki Hamada [a,1], Shigeyuki Oba [a], Shin Ishii [a,b]

[a] Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan
[b] ATR Cognitive Mechanisms Laboratories, Kyoto 619-0288, Japan

ARTICLE INFO

*Article history:*
Received 7 September 2017
Received in revised form 5 April 2018
Accepted 18 April 2018

*Keywords:*
Electroencephalography
Brain activity
Visual saliency
Decoding model

ABSTRACT

Understanding the functions of the visual system has been one of the major targets in neuroscience for many years. However, the relation between spontaneous brain activities and visual saliency in natural stimuli has yet to be elucidated. In this study, we developed an optimized machine learning-based decoding model to explore the possible relationships between the electroencephalography (EEG) characteristics and visual saliency. The optimal features were extracted from the EEG signals and saliency map which was computed according to an unsupervised saliency model (Tavakoli and Laaksonen, 2017). Subsequently, various unsupervised feature selection/extraction techniques were examined using different supervised regression models. The robustness of the presented model was fully verified by means of ten-fold or nested cross validation procedure, and promising results were achieved in the reconstruction of saliency features based on the selected EEG characteristics. Through the successful demonstration of using EEG characteristics to predict the real-time saliency distribution in natural videos, we suggest the feasibility of quantifying visual content through measuring brain activities (EEG signals) in real environments, which would facilitate the understanding of cortical involvement in the processing of natural visual stimuli and application developments motivated by human visual processing.



## 1. Introduction

Currently, there is still not enough understanding of how the human brain perceives the information input from the outside world and how neurons react correspondingly in a conscious and/or unconscious manner of perception. Non-invasive brain activity recording technologies, like electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), have been widely introduced to record human brain dynamics under certain circumstances. EEG measures brain activities over time with a high temporal resolution of milliseconds, while fMRI mainly identifies which area of the brain is in use with a high spatial resolution of millimeters. There is a convergent evidence that suggests that brain activity recordings play a vital role to boost the development of new biometric technologies and precede studies of brain functions like attention and memory (Han et al., 2015), motor control (Heimann, Umiltà, Guerra, & Gallese, 2014), and emotions (Alarcao & Fonseca, 2017).

In the fields of brain encoding and decoding, there has been a number of studies over the past decades addressing one of the basic questions of how information is represented in the brain (Naselaris, Kay, Nishimoto, & Gallant, 2011). As a consequence, the connections between brain activities in the visual cortex and low-level visual features such as orientation (Haynes & Rees, 2005), color (Brouwer & Heeger, 2009), and position (Thirion et al., 2006) have been intensively studied. To measure the brain responses while watching natural images, Kay, Naselaris, Prenger, and Gallant (2008) proposed an fMRI-based decoding system to apply a receptive field-based model to represent individual fMRI voxels. They modeled a generation process of fMRI signals in particular in visual areas V1, V2 and V3, and further tested the performance in an application of image identification. A high identification accuracy was obtained from two participants, suggesting the feasibility to predict novel natural images by using the proposed general visual decoder. Following this previous study, Naselaris, Prenger, Kay, Oliver, and Gallant (2009) proposed a Bayesian framework to model fMRI signals, and were successful in reconstructing natural images based on fMRI. According to their method, individual voxels were modeled in two different approaches, namely the Gabor wavelet-based structural encoding model and semantic-based encoding model, thus characterizing the fMRI responses in the early visual areas and anterior visual areas, respectively. Instead

* Corresponding author.
*E-mail addresses:* jane-l@sys.i.kyoto-u.ac.jp (Z. Liang), yasuyuki1004hamada@gmail.com (Y. Hamada), oba@i.kyoto-u.ac.jp (S. Oba), ishii@i.kyoto-u.ac.jp (S. Ishii).

[1] Yasuyuki Hamada is now affiliated with Nissan Motor Corporation.

https://doi.org/10.1016/j.neunet.2018.04.013
0893-6080/

of using static images as visual stimuli, Nishimoto et al. (2011) presented a motion energy-based encoding model to represent the fMRI signal patterns in the early visual areas while watching natural movies, and demonstrated the validity and temporal specificity of this encoding model. Furthermore, they applied a Bayesian decoder to test the reconstruction accuracy of using brain activity measurements to reconstruct the dynamic visual information in movies. Similarly, Han, Zhao, Hu, Guo, and Liu (2014) introduced an fMRI-based encoding model to predict the brain network response while the participants were free-viewing video clips. They pointed out that a successful brain encoding technique could benefit the evaluation and guidance of visual feature extraction in applications of visual attention and image processing, and could also boost the development of cognitive neuroscience studies.

In the abovementioned studies, however, researchers focused on human fMRI responses. To investigate the relationships between brain activities in a more natural environment and using natural visual stimuli such as natural images, Ghebreab, Scholte, Lamme, and Smeulders (2010) collected EEG signals from 32 participants while they were watching 700 natural scenes. They obtained a comparable identification accuracy to the previous study by Kay et al. (2008), and concluded that it is possible to predict natural images by using EEG responses as well. Nevertheless, quite a few studies have demonstrated that EEG features are effective in modeling brain activities upon perceiving natural visual stimuli. On the other hand, the band power oscillation in EEG recordings is one of the most important EEG features. From the analysis of oscillatory EEG components, the changes in EEG band powers in different frequency bands have been extracted and further employed to reveal a certain biological significance of the brain rhythmic oscillations (Klimesch, Schimke, & Schwaiger, 1994; Ray & Cole, 1985). The performance of EEG band powers has been broadly verified in many research dimensions such as brain memory system (Friese et al., 2013; Kawasaki, Kitajo, & Yamguchi, 2014), complex cognitive functions (Cohen, 2017; Fink & Benedek, 2014), emotions (Jenke, Peer, & Buss, 2014), motor system (Kajihara et al., 2015), and various applications in brain computer interface (Aliakbaryhosseinabadi, Kamavuako, Jiang, Farina, & Mrachacz-Kersting, 2017; Thomas & Vinod, 2016).

Furthermore, psychologists and physiologists have found that, while watching visual scenes, human beings tend to select the most important and informative portions from the visual scenes and conduct further analysis and understanding on the *selected* portions instead of the whole visual scenes (Koch & Ullman, 1985; Parasuraman, 1998). This kind of visual selective procedure is known as *visual attention*. With an interest in the mechanisms in early selective visual attention, the concept of *saliency map* was first proposed by Koch and Ullman (1985) to represent the conspicuity of a location in a visual scene and stand out how different this location is from its surroundings in terms of early representation features (e.g., color and orientation). Following the Koch and Ullman's idea, Itti et al. introduced the concept of *saliency map* in a manner of computational model and applied it to solve complex scene understanding problems (Itti, Koch, & Niebur, 1998). The efficiency of this saliency-driven approach has been verified in many studies. Based on this model, many computational models for predicting the image/video saliency have flourished (some details of background will be introduced in Section 2). More recently, Tavakoli and Laaksonen (2017) proposed an unsupervised learning-based saliency model, which is a more generic system for saliency estimation, as it does not require huge amount of training data like supervised learning-based models, and then it would not likely overfit to a specific database. As this algorithm was built based on unsupervised hierarchical features, we name it as UHF in this article. UHF utilized a hierarchical model based on Independent Subspace Analysis (ISA) with a hierarchy of features using natural image statistics, and benchmarked on two popular databases, MIT1003 and MIT300. It was found that UHF outperformed the existing popular bottom-up saliency-based models. More details about the UHF model are presented in Section 3.3.1. In the obtained saliency map, visual conspicuousness was well represented based on low-level feature contrast. Image pixels with high saliency values would carry important information in the image and can be processed further in a later stage of visual hierarchy in many real-world scenarios like object detection, recognition, and retrieval. Furthermore, the saliency map could be treated as an indicator to reflect the complexity of visual contents, i.e., the extent of involvement of important information in the input image.

As such, visual saliency has become a hot research topic in the past decades. Compared to the low-level visual features, e.g., orientation, spatial frequency, and color, visual saliency has been demonstrated to be a powerful and efficient representation of visual contents, thus promoting visual attention (Sharma, Jurie, & Schmid, 2012). In our current study, to further explore the relationship between the EEG features and visual stimuli contents, we focused on the best match of band power oscillations in EEG recordings and saliency features involved in video stimuli, and subsequently attempted to build an effective and robust decoding model to estimate the visual saliency in real-time based on the EEG features. The whole flowchart of the presented decoding pipeline is illustrated in Fig. 1. First, we recorded the human participants' EEG signals while they were watching video clips. Second, the EEG features in different frequency bands were extracted, and the visual saliency was subsequently computed at every video frame in terms of UHF. Third, feature selection/extraction techniques were implemented to reduce the EEG feature dimensionality, and the optimal EEG feature sets were sought. Finally, a computational regression model that associates EEG characteristics with saliency features was presented. The performance on the estimation of visual saliency has been fully demonstrated in the 10-fold or nested cross validation procedure, and the feasibility of usage of EEG signals for revealing the visual content has been discussed.

## 2. Computational models of visual attention

Inspired by functions of human visual system (HVS), computational models of the visual attention have undergone an explosive growth over the past two decades. In order to efficiently and effectively identify the regions/portions that are more important for HVS induced by various types of images and videos, attention/saliency detection problems have been tackled in different ways and can be summarized as below.

The past psychophysical studies suggested that when processing, locating and recognizing objects in the visual field, two major processes are involved: pre-attentive (bottom-up & primitive feature driven) and attentive (top-down & task driven) (Neisser, 1967). On the basis of this knowledge, existing computational models of the visual attention could be broadly grouped into three categories.

**(1) bottom-up based models**: being inspired by the bottom-up visual attention mechanism, which is a fast, automatic process triggered by low-level visual properties such as color, intensity, and orientation. The saliency values are determined by the primitive features in the visual stimuli. One of the most famous bottom-up based models is Itti et al.'s model (Itti et al., 1998) that proposed to model visual saliency in a topographical structure, by considering center–surround contrasts in terms of color, intensity, and orientation. Based on the input image, a saliency map was generated, in which high saliency values reflected high center–surround contrasts, e.g., the boundary of an object in the given image. In light of this, a number of studies have attempted to apply more effective methodologies and further improve the performance for saliency detection in both image and video stimuli.