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/© 2018 Elsevier Ltd. All rights reserved.

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.

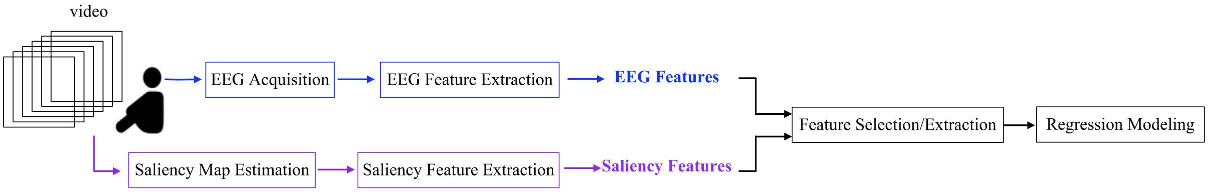


Fig. 1. Flowchart of the presented decoding pipeline.

For example, based on color channels, Zhang and Sclaroff (2013) proposed to measure the surroundedness cues (region contour information) using a set of Boolean maps for enforcing the figure-ground segregation and maximizing to segment the salient objects from the background. This model was reported to perform efficient detection of the saliency map, which was well demonstrated by using five benchmark eye tracking databases. In Mauthner et al.'s saliency model, the estimation of joint distribution of color and motion features was proposed (Mauthner, Possegger, Waltner, & Bischof, 2015). This study, based on the Gestalt theory, applied an integral histogram encoding vector to describe the local and global foreground saliency likelihoods and adaptively formed a final saliency map based on the individual saliency likelihoods.

(2) top-down based models: being inspired by the top-down visual attention mechanism, which is a high-level process involving the intention and thoughts and is guided by prior knowledge and given tasks. The saliency computation is determined by the given task. Consider an image including one apple and one orange, for example. In the bottom-up context, the apple and orange will be in the same saliency level. In the top-down modeling, however, orange may be assigned with a higher saliency value if the given task is to search the image for the orange. Top-down based saliency computing approaches have often been conducted through pre-defining the discriminant features and contextual guidance (Xu, Jiang, Wang, Kankanhalli, & Zhao, 2014), assigning previously learned weights onto different features (Zhao & Koch, 2011), adapting feature space in a supervised manner (Yang & Yang, 2017), and so on.

(3) a combination of bottom-up and top-down models: trying to integrate the two attention models above and introduce both of the low-level and high-level features to specific applications. In this type of saliency modeling, the bottom-up process plays a vital role in the detection of possibly salient regions from the input, while the top-down process leverages prior knowledge to pick up the salient parts that satisfy the requirement from the given task. Saliency computing models based on supervised learning-based models, such as regression (Zhou, Liu, Sun, Ye, & Wang, 2016), support vector machine (SVM) (Zhong, Liu, Liu, & Chung, 2010), and neural networks (NN) (Duan, Hu, Sun, & Duan, 2016), have been proposed. Especially after the development of deep learning architectures, many deep neural network (DNN)-based saliency computing methods have also been developed (John, Yoneda, Liu, & Mita, 2015; Kloss et al., 2016), and high performances have been achieved.

Considering the analysis approaches of the visual stimuli, on the other hand, the attention models can be also categorized into two types: **(1) feature-based models** and **(2) spatial-based models**. In the feature-based models, the saliency is only determined by the feature discriminant. Features are extracted from every pixel in the input, and saliency is computed in the pixel level as well, like in the Itti et al.'s model. However, psychophysical and neurophysiological researches evidenced the spatial structure-based attributes also affect the human attention (Foulsham & Underwood, 2008). So, the estimated saliency should not only be determined by the features in pixels, but also be affected by the structural organizations of the input. In the spatial-based models, the input is treated as patches/contiguous regions or perceptual objects, features are extracted from every patch/region/object, and the saliency computing is conducted on those spatial structures. The saliency model used in this study, i.e., UHF (unsupervised hierarchical features), belongs to the category of spatial-based attention models.

3. Materials and methods

3.1. Experimental procedures

3.1.1. Participants and stimuli

We recruited five male participants from our university (age, 22–25 years). All participants were right-handed, with normal or corrected-to-normal vision. They were all naive to visual experiments, have no history of neurological/eye diseases or disorders, and were not in use of any recreational drugs or other medicines. All experiments were approved by the Ethics Committee of the Graduate School of Informatics, Kyoto University (KUIS-EAR-2016-010). A signed consent form was obtained from every participant after explaining the aim and flow of the experiment.

A total of 10 video clips (video length, 58s each) were used in the experiment. All the video clips were taken from Activity Net (http://activity-net.org), covering different kinds of complex human activities in the daily living such as playing sports, cleaning up kitchen and food, caring animals, dancing and so on. The selected videos[2] were natural-colored, and there was no overlap in their contents. During the data acquisition, the videos were displayed on a standard 22-inch DELL LCD monitor. Participants were seated comfortably at a fixed viewing distance of 120 cm, with a corresponding subtended visual angle of stimuli presentation of $22° \times 14°$. To eliminate the auditory effects, all the videos were presented with the sound muted.

3.1.2. EEG data acquisition

Fig. 2 depicts a single experimental run including 10 trials in total. In each trial, after an eye fixation period of 4s (a white fixation cross was displayed in the center of a black screen), the participants were instructed to watch a video clip in a free-viewing mode (video viewing period), with eye movements allowed, for 58s. All the video clips were played in a randomized order to allow counter-balance between the participants. To minimize possible artifacts in the recorded EEG signals induced by head or body movements, a chin-rest was used to immobilize the participant's head.

While watching the video, EEG signals were simultaneously collected by using the Biosemi Active Two system, with 64 Ag/AgCI electrodes placed according to the standard international 10–20 electrode system (Jasper, 1958). The EEG signals from each electrode were digitized at a sampling rate (*fs*) of 256 Hz, with an amplifier bandpass of 0.16 to 100 Hz.

[2] Available on the project page https://sites.google.com/site/janezhenliang/eegsal.

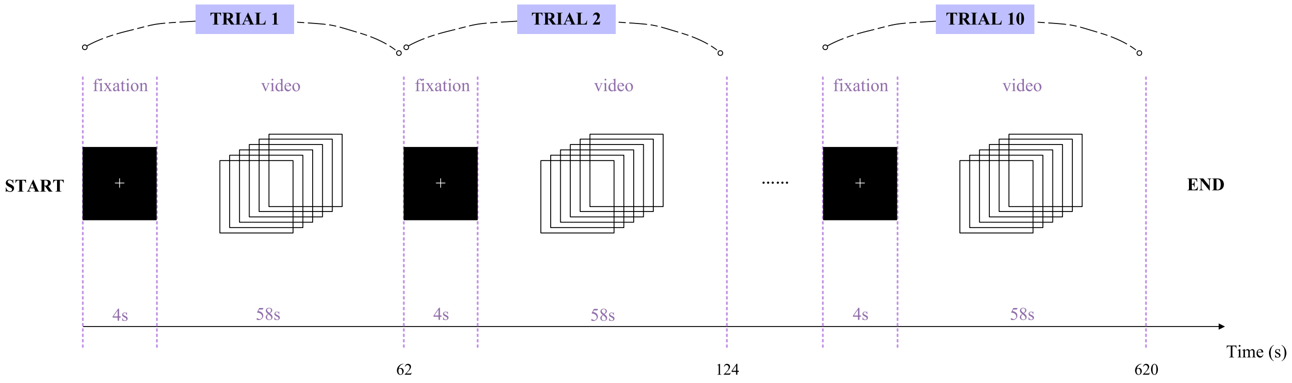


Fig. 2. Data acquisition procedure.

Table 1
Definition of the EEG frequency bands.

	Delta δ	Theta θ	Alpha α	Beta β	Gamma γ
Frequency range (Hz)	1–3	4–7	8–12	13–30	31–45

3.2. EEG data analysis

3.2.1. EEG signal preprocessing and channel selection

An FIR bandpass filter [0.5 Hz, 50 Hz] was applied to preprocess all the collected EEG signals. A common average reference was exploited by calculating the mean signal from all the 64 EEG channels, and this mean was subsequently extracted from the signals of each channel. Subsequently, baseline removal was conducted in each channel by subtracting the corresponding mean.

To focus on the EEG dynamics induced by visual saliency in the visual stimuli, we selected the channels located near the visual cortices. It is well established that the parietal lobe plays a critical role in the integration of sensory information in the human vision, while the occipital lobe is the visual processing center that contains most brain regions involved in vision (e.g., V1, V2, V3, V4, and V5). Therefore, based on three channel hemispheres (left, midline, and right hemispheres) and two lobes (parietal lobe and occipital lobe), a total of 19 channel locations were chosen: P1, P3, P5, P7, P9, P2, P4, P6, P8, P10, Pz, PO3, PO7, PO4, PO8, POz, O1, O2, and Oz.

3.2.2. EEG feature extraction

Here, the collected EEG data are denoted as follows: $\mathcal{D}^{fixation}_{c,s,t}$ (collected in the eye fixation period) and $\mathcal{D}^{video}_{c,s,t}$ (collected in the video viewing period), where c, s, and t correspond to a single selected EEG channel, a single participant, and a single trial, respectively. In particular, we have $c \in [1, 19]$, $s \in [1, 5]$, and $t \in [1, 10]$.

Referring to the past EEG studies (Finelli, Achermann, & Borbély, 2001; Gruber, Müller, Keil, & Elbert, 1999; Lin et al., 2010; Michels, Moazami-Goudarzi, Jeanmonod, & Sarnthein, 2008), we processed and normalized the collected EEG signals as follows. To extract the real-time EEG features, we applied a moving window of 2s to $\mathcal{D}^{fixation}_{c,s,t}$ with a sliding window step of 0.5s. This setting of the window size at 2s is fairly popular when analyzing human EEGs (Chu, Leahy, Pathmanathan, Kramer, & Cash, 2014; Gonuguntla & Veluvolu, 2017). A moving window of 2s could guarantee to capture the rich information in the frequency domain and maintain more details in the real-time changes. Thus, the EEG signals during the eye fixation period over 1024 points ($4s \times fs$) were divided into five sub-sequences. Similarly, we also applied a moving window of 2s to $\mathcal{D}^{video}_{c,s,t}$ with a sliding window step of 0.5s, in such a way that the signals in the video viewing period over 14,848 points ($58s \times fs$) were divided into 113 sub-sequences.

Power spectral density estimation algorithm (Welch, 1967) was applied to compute the spectral power distribution for each EEG sub-sequence (within a 2s moving window), using a Hamming window with 50% overlap. Subsequently, the average EEG power spectra in five typical EEG frequency bands (defined in Table 1) were extracted from each EEG channel. Thus, the EEG signals within a single moving window were characterized by a 95-dimensional feature vector (19 channels × 5 frequency bands), denoted as $\boldsymbol{F}_w$, where w indexes a single moving window.

Because there was no temporally changing stimuli during the eye fixation period, the EEG signals in that period were assumed stationary, and were treated as a baseline/reference for examining possible dynamics in the EEG signals during the subsequent video viewing period. More specifically, we averaged the EEG features over the eye fixation period using the following equation:

$$\overline{\boldsymbol{F}} = \frac{\sum_{w=1}^{n} \boldsymbol{F}_w^{fixation}}{n}, \tag{1}$$

where n was equal to 5, i.e., the number of moving windows in the eye fixation period. Subsequently, for characterizing the EEG signals induced by the video content only, the EEG features extracted in the video viewing period were normalized by the baseline as follows:

$$\mathcal{N}_w = \frac{\boldsymbol{F}_w^{video} - \overline{\boldsymbol{F}}}{\overline{\boldsymbol{F}}}, \qquad w = 1, \ldots, 113. \tag{2}$$

In order to build a subject-independent mapping between the EEG dynamics and video saliency, the EEG features obtained in Eq. (2) were, for each video, further normalized across the participants using the following equation:

$$\overline{\mathcal{N}}_w = \frac{\sum_{i=1}^{N} \mathcal{N}_w(i)}{N}, \qquad w = 1, \ldots, 113, \tag{3}$$

where N is the number of participants ($N = 5$). Note here that the averaged EEG features were specific to each video, although the video presenting order was different between the participants. Fig. 3 illustrates the steps followed to extract the EEG features from one experimental trial. Finally, for each trial, we obtained an EEG feature vector of 113 × 95 (spanning over a total of 113 moving windows).

3.3. Video saliency feature extraction

3.3.1. UHF model

Here, we introduce the background of UHF model (Tavakoli & Laaksonen, 2017) which was used to estimate saliency maps of

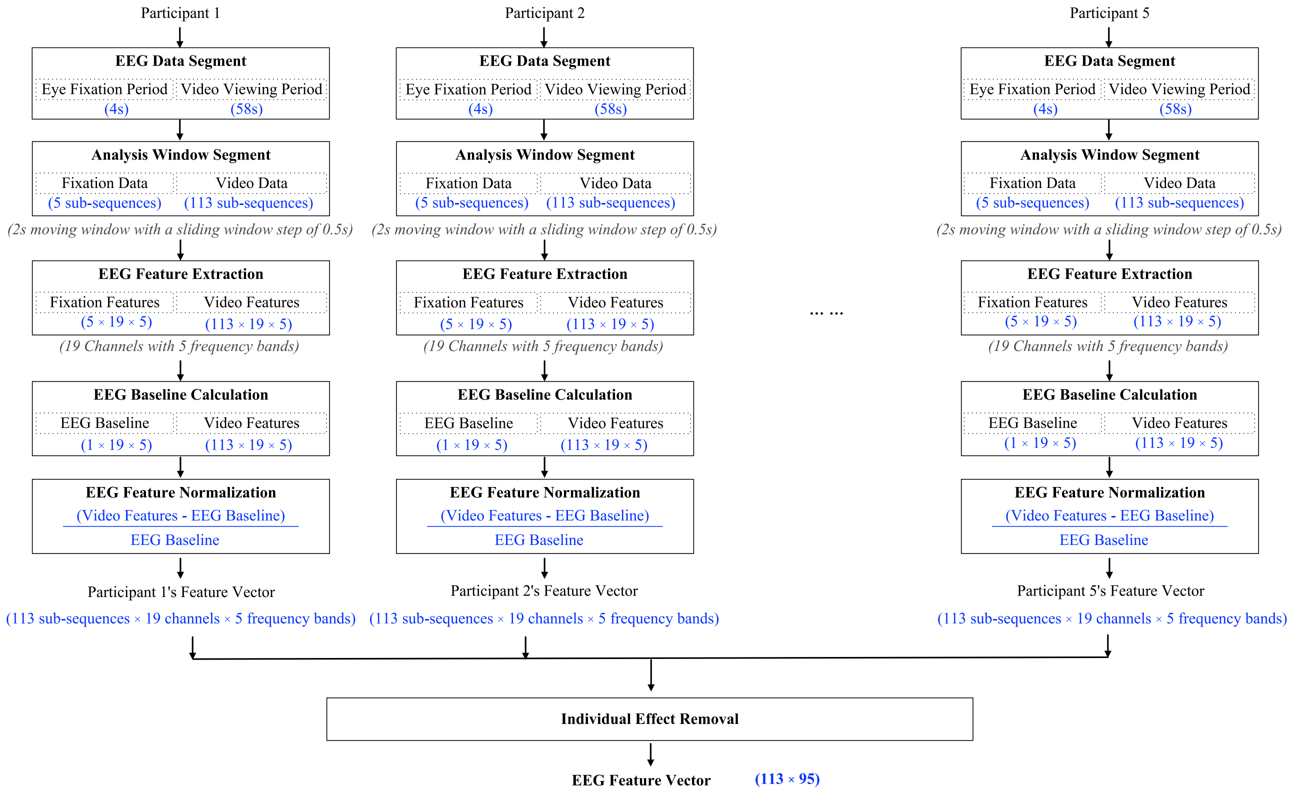


Fig. 3. The procedure used to extract the EEG features.

video frames. UHF, proposed by Tavakoli and Laaksonen (2017), is a saliency estimation model using stacked layers of Independent Subspace Analysis (ISA) in the spatial domain. In this model, hierarchical features were extracted in an unsupervised manner, and ISA filters were trained through a designed hierarchical architecture which performed convolution operations in a feed-forward manner, similar to deep neural networks. This model was originally used for multi-scale feature learning of natural image statistics. Fig. 4 shows the proposed flowchart.

As for the hierarchical features, both local saliency and global saliency were extracted from every image patch at every scale. The local saliency feature S_l was extracted from an image patch as,

$$S_l = -\sum_{i=1}^{n} \log(P(f_i)), \tag{4}$$

where f_i was the ith element of the n-dimension feature vector $\boldsymbol{f}$ in the image patch. The global feature was given as

$$S_g = \exp(-\frac{1}{k\boldsymbol{\pi}}), \tag{5}$$

where $\boldsymbol{\pi}$ was the equilibrium probability of a stochastic transition matrix $\boldsymbol{P} = \{p_{ij}\}$ and k was a smoothing factor. As shown in Eq. (6), p_{ij} was defined by feature distance, indicating the probability of transition from state/pixel i to j,

$$\begin{cases} p_{ij} = \dfrac{\exp(-\mathrm{D}(\boldsymbol{f}_i, \boldsymbol{f}_j))}{\sum_z \exp(-\mathrm{D}(\boldsymbol{f}_i, \boldsymbol{f}_z))} \\ \mathrm{D}\,(x, y) = \displaystyle\sum_{k=1}^{d} \frac{(x_k - y_k)^2}{x_k + y_k}. \end{cases} \tag{6}$$

Based on the extracted local and global features at every scale, the final saliency map was obtained as

$$S = \sum_{\sigma=1}^{n} \mathcal{N}(S_l^{\sigma} \times S_g^{\sigma}), \tag{7}$$

where $\mathcal{N}(\cdot)$ was a normalization using soft-max. In the experiment, these hierarchical features were extracted from the image patches in McGill Calibrated Color Image Database (Olmos & Kingdom, 2004), and the stacked ISA model with four scales ($n = 4$) was trained correspondingly.

3.3.2. Saliency feature extraction

Next, saliency features were extracted as Fig. 5. First, the saliency map for each video frame was obtained by using UHF, whereby the spatial resolution was the same as that of a single image. Second, the resolution of all the saliency maps was downsized (18 × 32) to simply reduce the computation time. Third, to make the temporal resolution of the saliency maps consistent with that of the EEG features, we averaged the saliency maps within a single 2s moving window. Subsequently, a series of temporally smoothed saliency maps were obtained (termed as $\boldsymbol{S}_w$), whose number was the same as that of the EEG features ($w = 1, 2, ..113$).

To fully represent the statistical properties of the obtained $\boldsymbol{S}_w$, we introduced three kinds of feature descriptors that measure the saliency maps in different statistical manners, and further compared the corresponding performances in Section 4. First, to measure the saliency variability/dispersion in a saliency map, we extracted the interquartile range (IQR) of elements in each saliency map as follows:

$$P_w = \mathcal{Q}_3 - \mathcal{Q}_1, \tag{8}$$

where $\mathcal{Q}_3$ and $\mathcal{Q}_1$ are the upper quartile (75th) and the lower quartile (25th) of the elements in $\boldsymbol{S}_w$, respectively. IQR is a statistic used to describe the mid-spread range, and one of the most useful dispersion measures (Hogg, McKean, & Craig, 2013). With P_w, we could observe the majority range of saliency values in a single saliency map, which is insensitive to possible outliers. Second, to measure the general tendency of saliency values in a saliency map,

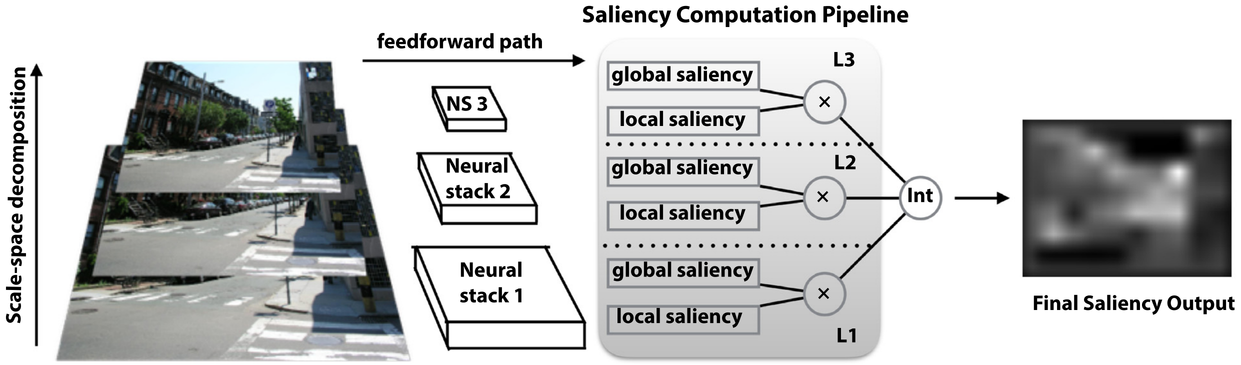


Fig. 4. The proposed flowchart in UHF (Tavakoli & Laaksonen, 2017).

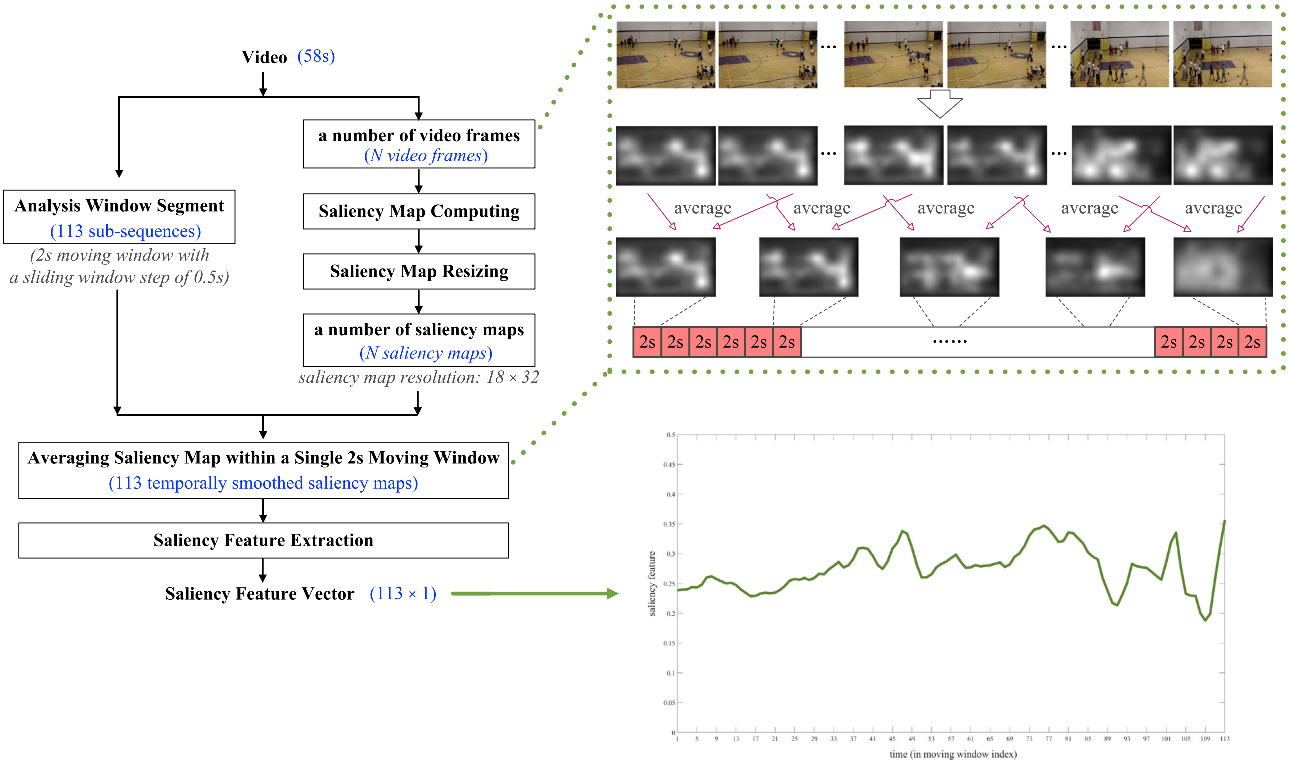


Fig. 5. The procedure to extract video saliency features.

we examined the mean or average as

$$A_w = \frac{\sum_{j=1}^{q} \sum_{k=1}^{l} \boldsymbol{S}_w}{q \times l}, \tag{9}$$

where q and l are the width and height of a single saliency map, respectively. More specifically, q corresponds to 18 and l corresponds to 32. Third, the maximum saliency value in a saliency map was also explored as

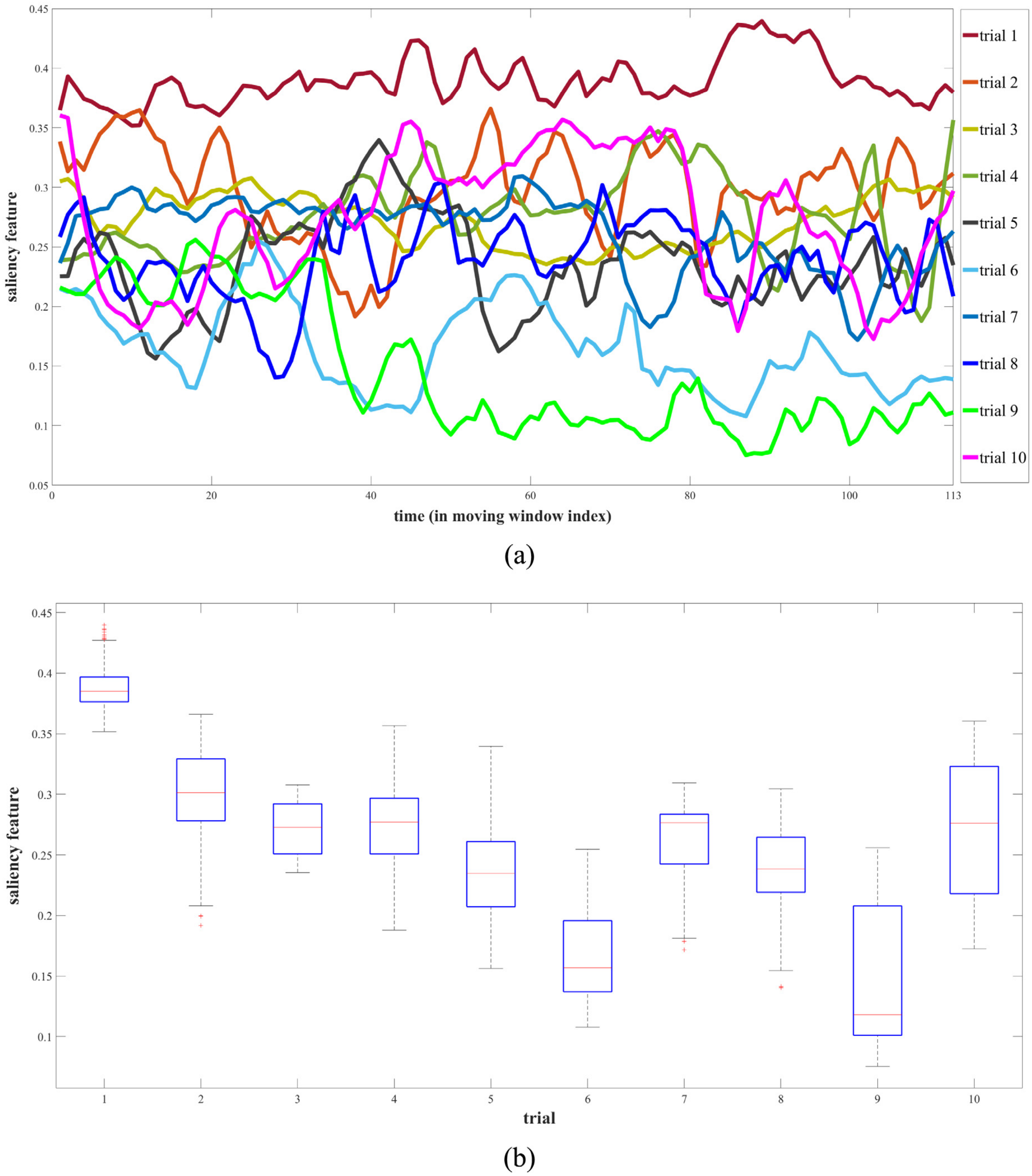
$$Z_w = \max_{j=1,\ldots,q} \max_{k=1,\ldots,l} \boldsymbol{S}_w. \tag{10}$$

Here, Z_w indicates the saliency value of the largest conspicuous pixel in the visual input.

Accordingly, for each trial, we had a series of 95-dimensional EEG features and a series of saliency features, along the 113 time-points. As shown in Fig. 6, it is observed that the saliency features (P_w) fluctuated over time even in a single trial, and varied between different trials. We found that the saliency feature in video 1 (trial 1) was dominantly in a high level, while the saliency feature behaved relatively lower in video 9 (trial 9), especially in the later part. We compared the video content and the corresponding saliency map to the obtained saliency feature (P_w) in videos 1 and 9 (Fig. 7), and observed that the extracted saliency feature does reflect the complexity of the video content.

3.4. Feature selection/extraction

Putting all the 10 trials together, we produced an EEG feature matrix $\boldsymbol{X}$ with a size of 1130 × 95 (single trial's EEG features: 113 × 95), and a saliency feature vector $\boldsymbol{Y}$ with a size of 1130 × 1 (single trial's saliency features: 113 × 1). In the following, we explored the possible relationships between EEG features $\boldsymbol{X}$ and saliency features $\boldsymbol{Y}$, and built an optimized regression model $\boldsymbol{Y} = f(\boldsymbol{X})$ to estimate the saliency feature based on the EEG features along the time. To avoid the curse of dimensionality and achieve a better modeling performance, before regression modeling, we first reduced the EEG feature dimension $\mathbb{R}^d (d = 95)$ to a lower dimension feature space $\mathbb{R}^g (g < d)$ by selecting or creating the



(a)

(b)

Fig. 6. Time-series of the extracted saliency features (P_w) for 10 videos (10 trials). (a) fluctuation of the saliency feature over time; (b) a boxplot to show the distribution of the saliency feature in each trial.

most discriminant/representative features. Two types of feature selection/extraction methods were applied.

(1) selecting a subset of features from the original feature space: two popular feature selection methods, namely, minimum redundancy maximum relevance feature selection (mRMR) (Peng, Long, & Ding, 2005) and recursive feature elimination (RFE) (Mahmoodian & Ebrahimian, 2016), were adopted. The main idea of the mRMR method is employing mutual information to optimize the feature set and to balance between the relevance and redundancy, while RFE tries to explore the optimal feature set by repeating removal of a single feature with a lower supervised weight.

(2) creating a new set of transformed features: performances with singular value decomposition (SVD) (Gentle, 1998), principal component analysis (PCA) (Jolliffe, 1986), and kernel principal component analysis (KPCA) (Schölkopf, Smola, & Müller, 1998) were tested. In SVD, $\boldsymbol{X}$ is decomposed as

$$\boldsymbol{X} = USV^T = [u_1 \dots u_d] \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_d \end{bmatrix} \begin{matrix} v_1 \\ [\,\vdots\,], \\ v_d \end{matrix}$$

$$\{s_1 \geq s_2 \geq \cdots \geq s_d\}\, \&\, \boldsymbol{X} \in \mathbb{R}^d \tag{11}$$

where U is the eigenvectors of $\boldsymbol{XX}^T$, S is the square root of the obtained eigenvalues of $\boldsymbol{XX}^T$, and V is the eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$. Only the top-$g$ largest singular values in the diagonal matrix S are kept to reconstruct an approximated matrix that retains as much

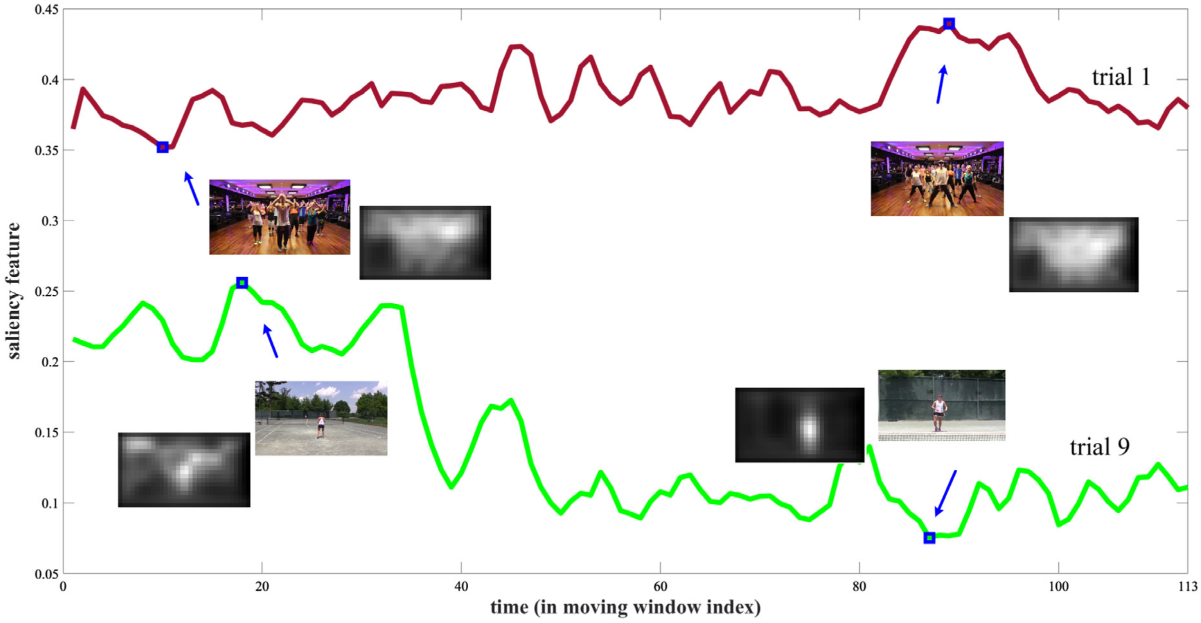


Fig. 7. Examples of video content with the corresponding saliency map on the time-series of the extracted saliency feature (P_w) for videos 1 and 9 (trials 1 and 9, respectively).

variance as possible from the original matrix, denoted as:

$$\boldsymbol{X}^g = \left[u_1 \ldots u_g\right] \begin{bmatrix} s_1 & & \\ & \ddots & \\ & & s_g \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_g \end{bmatrix}, \quad \boldsymbol{X}^g \in \mathbb{R}^g (g < d). \tag{12}$$

While the PCA attempts to project the data to a low-dimensional sub-space in which the maximum variance of the data distribution is retained. The computation procedure of the PCA is almost identical to the SVD except for that column vectors of matrix $\boldsymbol{X}$ are centralized before the application of the SVD, leading to a little different results. The KPCA is an extended version of PCA based on kernel methods (Schölkopf et al., 1998). Through kernel mapping, KPCA is capable of identifying a nonlinear sub-space.

3.5. Regression modeling

After feature selection/extraction, EEG features were presented to a lower feature dimension space, denoted as $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ with vector $\boldsymbol{x}_i \in \mathbb{R}^g$ (g is feature dimensionality, $g < d$, $d = 95$), and an associated set of saliency features, $\boldsymbol{Y} = \{y_1, \ldots, y_n\}$, with $y_i \in \mathbb{R}$. Here, n denotes the number of samples (here, feature vectors) in the dataset. To investigate the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$, three types of regression models were explored to build a function f that is capable of minimizing the discrepancy between the measured y_i and the model's prediction/reproduction $y_i' = f(\boldsymbol{x}_i)$.

The most popular regression technique on multiple variables is the multiple linear regression (MLR) (Freedman, 2009), in which the relationship between the input and output is linear. However, the linearity assumption can deteriorate the performance especially when there is underlying non-linearity. Considering the possible non-linearity between the saliency and EEG features, we utilized two kinds of nonlinear regression techniques, namely, the support vector regression (SVR) and kernel ridge regression (KRR). The SVR (Cortes & Vapnik, 1995) is a nonlinear regression technique which reveals superior modeling capabilities, and has been used in various regression and classification applications (Georga et al., 2013). Instead of the least-squared error loss function, SVR adopts an ε-insensitive loss function to obtain the solutions and hence achieves high generalization ability; it is defined as

$$\begin{aligned} y' &= \boldsymbol{w} \cdot \Phi(\boldsymbol{x}) + \xi \\ &= \sum_{i=1}^{n} (a_i - a_i^*) \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}) + \xi \\ &= \sum_{i=1}^{n} \left(a_i - a_i^*\right) K(\boldsymbol{x}_i, \boldsymbol{x}) + \xi, \end{aligned} \tag{13}$$

where $\boldsymbol{w}$ and ξ, which are both estimated, are the weight and bias, respectively. Here a_i and a_i^* are the optimal Lagrange multipliers determined based on the training dataset. $K(\boldsymbol{x}, \boldsymbol{z})$ is the Gaussian kernel defined as $\exp(-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2})$. On the other hand, the KRR (Murphy, 2012) is a linear regression in the nonlinear space, with an L2 regularization on the weight parameter, determined as

$$\operatorname*{argmin}_{\boldsymbol{w}} \frac{1}{2} \sum_{j=1}^{n} (y_j - \boldsymbol{w}^T \Phi(\boldsymbol{z}_j))^2 + \frac{1}{2} \lambda \|\boldsymbol{w}\|^2, \tag{14}$$

where λ is a positive constant for regularization. A solution of Eq. (14) is given by a regularized least square, as

$$\boldsymbol{w} = (\boldsymbol{G} + \lambda \mathbf{I})^{-1} \boldsymbol{Y}, \tag{15}$$

where $\boldsymbol{G} = \left[\left\langle \Phi(\boldsymbol{z}_i), \Phi(\boldsymbol{z}_j) \right\rangle\right]_{n \times n}$ is the Gram matrix. When a new test data point $\boldsymbol{x}$ comes, the corresponding y' can be calculated by projecting $\boldsymbol{x}$ onto the estimated $\boldsymbol{w}$ as follows:

$$y' = \boldsymbol{w}^T \Phi(\boldsymbol{x}). \tag{16}$$

4. Experimental results

4.1. Evaluation criteria

The results were evaluated in terms of two criteria: Pearson correlation coefficient (PCC) and normalized mean squared error (NMSE). PCC between the target $\boldsymbol{Y}$ and the regression result $\boldsymbol{Y}'$ is given by:

$$\rho\left(\boldsymbol{Y}, \boldsymbol{Y}'\right) = \frac{cov(\boldsymbol{Y}, \boldsymbol{Y}')}{\sigma_{\boldsymbol{Y}} \sigma_{\boldsymbol{Y}'}} = \frac{1}{n-1} \sum_{i=1}^{n} (\frac{y_i - \mu_{\boldsymbol{Y}}}{\sigma_{\boldsymbol{Y}}})(\frac{y_i' - \mu_{\boldsymbol{Y}'}}{\sigma_{\boldsymbol{Y}'}}), \tag{17}$$

where $\mu_{\boldsymbol{Y}}$ and $\mu_{\boldsymbol{Y}'}$ are the means of $\boldsymbol{Y}$ and $\boldsymbol{Y}'$, respectively, and $\sigma_{\boldsymbol{Y}}$ and $\sigma_{\boldsymbol{Y}'}$ are the standard deviations of $\boldsymbol{Y}$ and $\boldsymbol{Y}'$, respectively; $\rho\left(\boldsymbol{Y}, \boldsymbol{Y}'\right)$ ranges in $[-1,1]$. A positive/negative value indicates a positive/negative linear relationship, while a zero value suggests no correlation. As a reference measure, we calculated the *p*-value for the given PCC value, corresponding to the null-hypothesis that the correlation between $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ is zero under a two-sided test. On the other hand, NMSE denotes the normalized L2 difference between $\boldsymbol{Y}$ and $\boldsymbol{Y}'$, and is given by

$$\mathrm{NMSE} = 1 - \frac{\sum_{i=1}^{n}\left(y_i' - y_i\right)^2}{\sum_{i=1}^{n}\left(y_i' - \overline{y'}\right)^2}, \tag{18}$$

where $\overline{y'} = \frac{1}{n}\sum_{i=1}^{n} y_i'$. Note that NMSE is related to the objective function of the regression models but normalized by the variance. NMSE ranges between –Inf (worst) and 1 (best). It is expected to be close to 1 if the regression model is optimized well to perform the prediction in both of the space and time domains.

4.2. Modeling results

4.2.1. Results with feature selection/extraction

In this study, we first performed saliency estimation by using the feature selection approaches: mRMR and RFE. The 10-fold cross validation procedure was run for each combination of feature selection and regression algorithm, and the regression performances on the test (validation) dataset were examined (Fig. 8). The feature dimension g was varied from 5 to 95 with a step of 5. Such cross-validation procedure showed that the best PCC was 0.89 with p-value$<<$0.001, when mRMR+KRR was applied to the prediction of visual saliency P_w under $g = 35$. Meanwhile, the best NMSE result was 0.72 when g was equal to 40. Comparing to MLR, modeling with SVR and KRR showed better performance. It was thus evident that non-linear modeling would be more suitable to solve this decoding problem. Also, P_w showed better performance in general and could be considered as a more effective feature representation for visual saliency in natural videos and a better target to be predicted based on EEG features.

Further, evaluations using SVR and KRR regression algorithms with feature extraction approaches were performed for the prediction of P_w. The regression performance was tested when EEG feature dimensionality was reduced by SVD, PCA, or KPCA, respectively. The corresponding 10-fold cross validation results on the test dataset were reported in Table 2, and different percentages of variance retained, after the feature extraction, were concerned. Note here that the *p*-value was calculated for each null hypothesis that the corresponding correlation coefficient is zero without multiplicity correction.

The regression performance highlighted that KPCA performed better than SVD or PCA. Here, KPCA was tuned with different kernel functions (linear, polynomial, and Gaussian) and different kernel parameter settings (various polynomial orders and Gaussian kernel's σ values). It was found when a linear or polynomial kernel was applied, the decoding performance was very poor. If KPCA was working with a Gaussian kernel, on the other hand, much better results could be obtained; moreover, the performance was not sensitive to the σ value especially when σ was in the range from 2 to 3. We found the best CV performance was obtained by KPCA+SVR: the best PCC value was 0.91 when KPCA's σ was set to 2; the best NMSE result was 0.73 when KPCA's σ was equal to 2.5. These results evidenced that, in this decoding application, mapping EEG features onto a low-dimensional transformed feature space tended to outperform selecting a subset of features from the original set. Next, we further tried to tune the hyperparameters in KPCA+SVR and KPCA+KRR, with 95% variance retained.

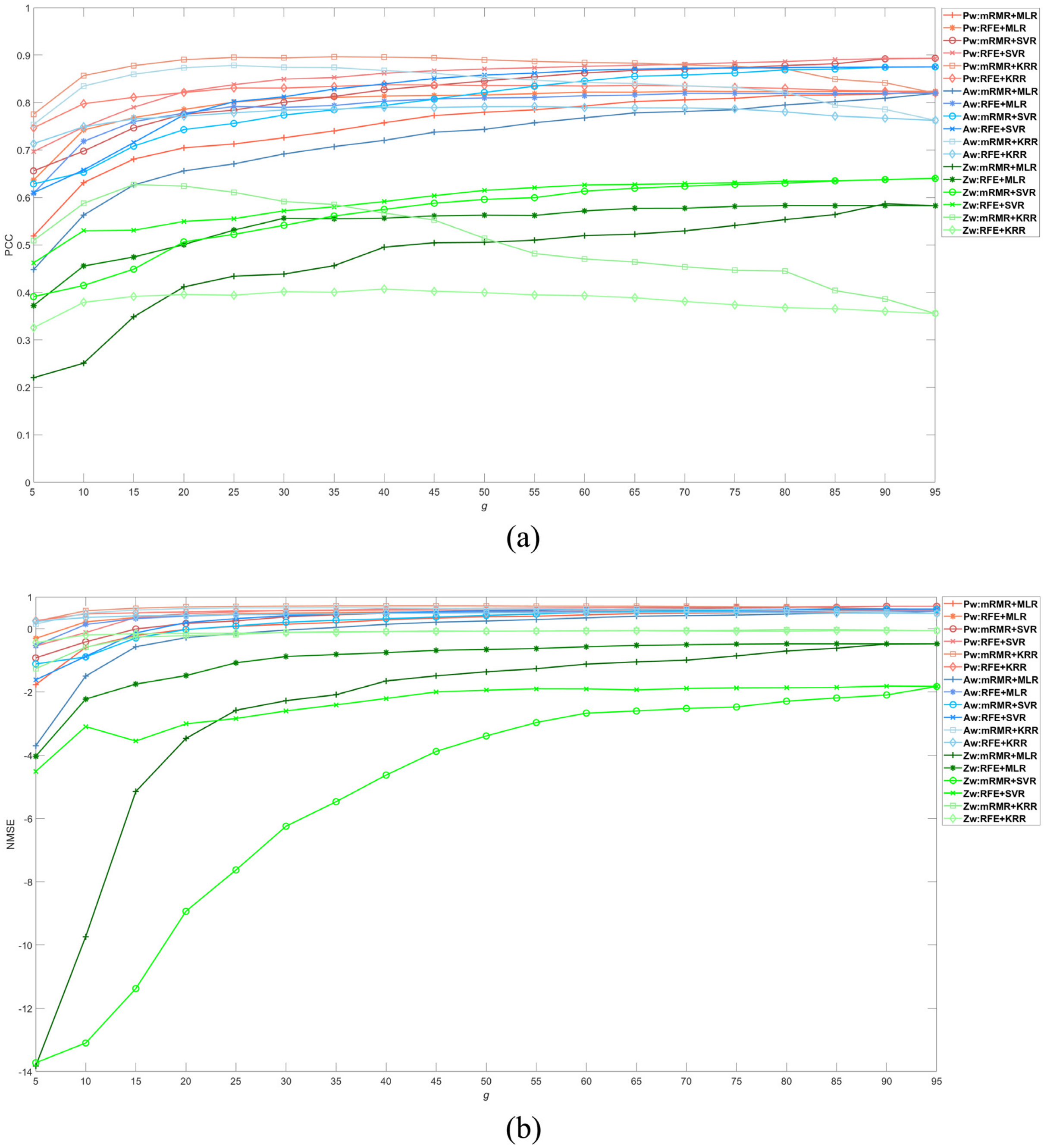
4.2.2. Parameter tuning in KPCA+SVR and KPCA+KRR

We utilized the nested cross validation (CV) here rather than the ordinary CV (e.g. 10-fold CV) to estimate the optimized hyperparameters' performances; the ordinary CV may overestimate the performance when taking the best result among a number of validation trials due to statistical variance, whereas the nested CV can reduce the bias and estimate the variance (Stone, 1978). In the nested CV of KPCA+SVR, SVR's kernel type was adjusted to linear, polynomial and Gaussian; the polynomial order together with the polynomial kernel and the σ value together with the Gaussian kernel were optimized in the range of $\{2, 3, 4\}$ and $\{0.5, 1, 1.5, 2, 2.5, 3,$ automatic hyperparameter tuning in MATLAB's 'fitrsvm' function$\}$, respectively. In the results, polynomial kernel with the polynomial order of 2 was consistently selected through the *inner loop*, and the model with the selected hyperparameter was tested in the *outer loop*. As the consequence, the estimated PCC was 0.92 and the estimated NMSE was 0.84. When the KPCA's σ value was changed to 2.5 or 3, slightly lower PCC and NMSE were obtained ($\sigma = 2.5$: PCC: 0.91, NMSE: 0.82; $\sigma = 3$: PCC: 0.89, NMSE: 0.77).

The hyperparameters in KRR were also evaluated and optimized in the nested CV procedure, in which σ value in the Gaussian kernel was varied over $\{0.5, 1, 1.5, 2, 2.5, 3\}$ and λ value was adjusted in the range of $[10^{-7}, 10^{4}]$. The nested CV results reported. $\sigma = 3$ was consistently selected and the modeling performance was not sensitive to the regularization constant if λ was in a range between 10^{-7} and 10^{-3}. The estimated value of optimized PCC was 0.94 and that of NMSE was 0.85. Similarly, the KPCA+KRR's performance was also tested when the KPCA's σ value was equal to 2.5 or 3, almost the same PCC and NMSE were obtained, which again evidenced the performance of KPCA+KRR was not sensitive to the KPCA's σ value when σ was in the range from 2 to 3. Accordingly, KPCA+KRR with the optimized hyperparameters performed the best in this application. When compared with KPCA+SVR (PCC: 0.92, NMSE: 0.84), a higher PCC value (0.94) and NMSE value (0.85) were obtained in the nested CV procedure. We termed this optimized decoding pipeline as **KPCA+KRR** hereafter, in which KPCA employed Gaussian kernel function with the σ value of 2 and KRR used Gaussian kernel function with the σ value of 3 and the regularization constant λ of 10^{-5}.

4.2.3. Modeling on the other visual saliency models

This study attempted to present a decoding pipeline of EEG characteristics for prediction/estimation of visual saliency in videos. Although the UHF model was used as an example to extract the saliency information from videos in this study, this was to prove the basic idea of decodability. In other words, the developed **KPCA+KRR** model should be adaptive to other visual saliency models which can correctly measure the saliency values in the visual content. To demonstrate this adaptivity of **KPCA+KRR**, we further tested the presented decoding pipeline on the saliency results obtained with other popular visual saliency models, e.g., the Itti et al.'s model. Note that, for each saliency model, we downloaded the source codes provided by the original authors, re-computed the saliency maps of all the videos, extracted the corresponding P_w saliency features as presented in Section 3.3.2, and further evaluated by the developed **KPCA+KRR** model in a CV manner. The corresponding performances were presented in Table 3. It was observed **KPCA+KRR** with various visual saliency models exhibited promising results, which evidenced the generality and adaptivity of the presented pipeline and proved a success in decodability of EEG characteristics to predict visual saliency embedded in the videos.



(a)

(b)

Fig. 8. 10-fold cross-validation performance of predicting visual saliency P_w, A_w, and Z_w using variety of feature size g of EEG characteristics. (a) PCC and (b) NMSE.

5. Discussions and conclusion

In the present study, we explored the possible relationship between the EEG activities and saliency embedded in visual stimuli, and developed an optimized decoding model that well predicted the visual saliency distribution based on the EEG characteristics.

In accordance with the results obtained in the Steady State Visual Evoked Potentials (SSVEP) study (Uribe et al., 2014), placements of a total of 19 EEG electrodes were first chosen and the corresponding frequency domain features were extracted. Subsequently, various machine learning models together with different feature selection/extraction methods were implemented to find the key EEG characteristics that are evoked, which thus would be involved in the visual attention and have a strong relationship with the saliency distribution. Finally, **KPCA+KRR**, an optimal decoding model under a combination of unsupervised feature extraction and supervised regression, was presented to estimate the changes of visual saliency based on the extracted EEG characteristics. Very promising prediction/reconstruction performance was achieved in the CV procedure and the generality and adaptivity of the present pipeline were also demonstrated. The best prediction performance was: PCC = 0.94 and NMSE = 0.85. We confirmed through a nested CV that this promising performance was minimally overestimated.

Considering to capture the rich frequency features and to maintain more details in EEG dynamics, we set the moving window size at 2s to perform continuous mapping between EEG signals and visual saliency. We believe the visual latency between the appearance of a video frame and the evoked response in the EEG signals could be neglected. One of the reasons is that the most responsible regions for visual saliency, occipital regions like V1 and

Table 2
10-fold cross-validation performance of predicting visual saliency P_w when SVD, PCA, or KPCA was used to reduce the EEG feature dimensionality.

P_w	Percentage of variance retained (%)							
	80		85		90		95	
	PCC	NMSE	PCC	NMSE	PCC	NMSE	PCC	NMSE
SVD								
SVR	0.77***	−0.87	0.80***	−0.70	0.81***	−0.56	0.81***	−0.41
KRR	0.89***	0.63	0.90***	0.63	0.90***	0.58	0.90***	0.52
PCA								
SVR	0.67***	−0.58	0.79***	0.35	0.81***	0.44	0.88***	0.68
KRR	0.69***	0.08	0.78***	0.43	0.78***	0.46	0.83***	0.59
KPCA (Gaussian, $\sigma = 0.5$)								
SVR	0.88***	0.68	0.88***	0.69	0.88***	0.69	0.88***	0.69
KRR	0.88***	0.66	0.88***	0.66	0.88***	0.66	0.88***	0.67
KPCA (Gaussian, $\sigma = 1$)								
SVR	0.90***	0.67	0.90***	0.68	0.90***	0.68	0.90***	0.68
KRR	0.87***	0.65	0.87***	0.65	0.87***	0.65	0.87***	0.65
KPCA (Gaussian, $\sigma = 1.5$)								
SVR	0.90***	0.69	0.90***	0.69	0.90***	0.69	0.90***	0.69
KRR	0.87***	0.61	0.87***	0.61	0.87***	0.61	0.87***	0.61
KPCA (Gaussian, $\sigma = 2$)								
SVR	**0.91*****	0.72	**0.91*****	0.72	**0.91*****	0.72	**0.91*****	0.72
KRR	0.89***	0.66	0.89***	0.66	0.89***	0.66	0.89***	0.66
KPCA (Gaussian, $\sigma = 2.5$)								
SVR	0.90***	**0.73**	0.90***	**0.73**	0.90***	**0.73**	0.90***	**0.73**
KRR	0.90***	0.70	0.90***	0.70	0.90***	0.70	0.90***	0.70
KPCA (Gaussian, $\sigma = 3$)								
SVR	0.90***	0.72	0.90***	0.72	0.90***	0.72	0.90***	0.72
KRR	0.90***	0.72	0.90***	0.72	0.90***	0.72	0.90***	0.72

**p*-value<0.05; ***p*-value<0.001; ****p*-value<<0.001.

Table 3
10-fold cross validation performance of **KPCA+KRR** when other popular visual saliency models were employed.

Percentage of variance retained (%)							
80		85		90		95	
PCC	NMSE	PCC	NMSE	PCC	NMSE	PCC	NMSE
GVBS (Harel, Koch, & Perona, 2007)							
0.89***	0.70	0.89***	0.70	0.89***	0.71	0.89***	0.71
Itti et al.'s model (Itti et al., 1998)							
0.95***	0.88	0.95***	0.88	0.95***	0.88	0.95***	0.88
Duan et al.'s model (Duan, Wu, Miao, Qing, & Fu, 2011)							
0.94**	0.87	0.95**	0.87	0.95**	0.87	0.95**	0.87
Hou et al.'s model (Hou, Harel, & Koch, 2012)							
0.95***	0.87	0.95***	0.87	0.95***	0.87	0.95***	0.87
UHF (Tavakoli & Laaksonen, 2017)							
0.94***	0.85	0.94***	0.85	0.94***	0.85	0.94***	0.85

**p*-value<0.05; ***p*-value<0.001; ****p*-value<<0.001.

V2/3, showed good agreement with existing neuroscience studies (de Graaf, Koivisto, Jacobs, & Sack, 2014; Emmanouil, Avigan, Persuh, & Ro, 2013; Koivisto, Mäntylä, & Silvanto, 2010). The latency from visual stimuli to be processed in such early/early-middle visual areas is known as 100∼200 ms, which is much shorter than the moving window size (2s) used in our decoding model. Also, if considering the most famous evoked EEG, P300, it still has 300 ms latency from the onset of visual stimuli. We actually slightly shifted the EEG signals with 200 ms and did some regression experiments, but the results were almost the same. So, the visual latency could be negligible in this study.

To the best of the authors' knowledge, a machine learning-based decoding model for using EEG features to predict saliency changes has not been previously presented. The major contributions of our current study are summarized as follows:

1. The study addressed the possible relationships between visual saliency and brain activity (EEG signals), and indicated positive correlations.
2. The study developed a real-time decoding model, by which the saliency distribution in videos could be accurately estimated using EEG characteristics.
3. The presented model offered a novel way for modeling brain encoding and decoding processing on the basis of EEG signals. Indeed, the visual saliency was suggested as a useful index to model brain activities in visual processing tasks.

4. The study provided an intuitive approach leading to EEG-based encoding/decoding studies in the visual attention domain. A successful EEG-based brain decoding model, in turn, could be used as a guidance to select valid visual features that better reflect the human processing of visual stimuli, and further benefit the development of more biologically-inspired visual attention and saliency models.

In summary, this study verified the possibility of using brain activities to predict features in visual contents and further presented a computational decoding model that well predicted/reconstructed the saliency distribution based on EEG characteristics. We searched for a set of optimal features that worked as better indicators for both saliency maps and EEG signals. The promising results were achieved and fully demonstrated in the cross validation procedure. Limitations of this study include the limited number of participants and video clips that were examined at the current stage. In the future, the number of participants will be increased, and a larger number of videos with various saliency distributions will be examined. For example, with the general concept of saliency computing, the selected video samples could be categorized into five types: natural videos without any obvious object, single dominant object with pure/simple background, single dominant object with complex background, multiple dominant objects with pure/simple background, and multiple dominant objects with complex background. Such categorization of the video samples may provide a better and deeper understanding of the relation of visual saliency and ongoing brain activity.

Acknowledgment

This study was supported by the Project for Next-Generation Core Robotic and AI Technology Development (New Energy and Industrial Technology Development Organization [NEDO]).

References

Alarcao, S. M., & Fonseca, M. J. (2017). Emotions recognition using EEG signals: a survey. *IEEE Transactions on Affective Computing*,(99), 1–20 (in print).

Aliakbaryhosseinabadi, S., Kamavuako, E. N., Jiang, N., Farina, D., & Mrachacz-Kersting, N. (2017). Classification of EEG signals to identify variations in attention during motor task execution. *Journal of Neuroscience Methods*, *284*(1), 27–34.

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *The Journal of Neuroscience*, *29*(44), 13992–14003.

Chu, C. J., Leahy, J., Pathmanathan, J., Kramer, M. A., & Cash, S. S. (2014). The maturation of cortical sleep rhythms and networks over early development. *Clinical Neurophysiology*, *125*, 1360–1370.

Cohen, M. X. (2017). Where does EEG come from and what does it mean? *Trends in Neurosciences*, *40*(4), 208–218.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

de Graaf, T. A., Koivisto, M., Jacobs, C., & Sack, A. T. (2014). The chronometry of visual perception: review of occipital TMS masking studies. *Neuroscience and Biobehavioral Reviews*, *45*, 295–304.

Duan, L., Wu, C., Miao, J., Qing, L., & Fu, Y. (2011). Visual saliency detection by spatially weighted dissimilarity. In *2011 IEEE conference on computer vision and pattern recognition* (pp. 473–480).

Duan, P., Hu, B., Sun, H., & Duan, Q. (2016). Saliency detection based on BP-neural network. In *2016 12th world congress on intelligent control and automation* (pp. 551–555).

Emmanouil, T. A., Avigan, P., Persuh, M., & Ro, T. (2013). Saliency affects feedforward more than feedback processing in early visual cortex. *Neuropsychologia*, *51*(8), 1497–1503.

Finelli, L. A., Achermann, P., & Borbély, A. A. (2001). Individual 'fingerprints' in human sleep EEG topography. *Neuropsychopharmacology*, *25*, S57–S62.

Fink, A., & Benedek, M. (2014). EEG alpha power and creative ideation. *Neuroscience & Biobehavioral Reviews*, *44*, 111–123.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2), 6, 1–17.

Freedman, D. A. (2009). *Statistical models: Theory and practice* (2nd ed.). Cambridge University Press.

Friese, U., Köster, M., Hassler, U., Martens, U., Trujillo-Barreto, N., & Gruber, T. (2013). Successful memory encoding is associated with increased cross-frequency coupling between frontal theta and posterior gamma oscillations in human scalp-recorded EEG. *NeuroImage*, *66*(1), 642–647.

Gentle, J. E. (1998). *Numerical linear algebra for applications in statistics*. New York, Berlin: Springer-Verlag.

Georga, E. I., Protopappas, V. C., Ardigo, D., Marina, M., Zavaroni, I., Polyzos, D., et al. (2013). Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression. *IEEE Journal of Biomedical and Health Informatics*, *17*(1), 71–81.

Ghebreab, S., Scholte, S., Lamme, V., & Smeulders, A. (2010). Rapid natural image identification based on EEG data and Global Scene Statistics [Abstract]. *Journal of Vision*, *10*(7), 1394.

Gonuguntla, V., & Veluvolu, K. C. (2017). Emotion associated brain function network analysis in human EEG using graph measures. In *Proceedings of academics world international conference* (pp. 11–15).

Gruber, T., Müller, M. M., Keil, A., & Elbert, T. (1999). Selective visual-spatial attention alters induced gamma band responses in the human EEG. *Clinical Neurophysiology*, *110*(12), 2074–2085.

Han, J., Chen, C., Shao, L., Hu, X., Han, J., & Liu, T. (2015). Learning computational models of video memorability from fMRI brain imaging. *IEEE Transactions on Cybernetics*, *45*(8), 1692–1703.

Han, J., Zhao, S., Hu, X., Guo, L., & Liu, T. (2014). Encoding brain network response to free viewing of videos. *Cognitive Neurodynamics*, *8*(5), 389–397.

Harel, J., Koch, C., & Perona, P. (2007). *Advances in neural information processing systems*: *Vol. 19*. *Graph-based visual saliency* (pp. 545–663). Cambridge, MA: MIT Press.

Haynes, J. D., & Rees, G. (2005). predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, *8*, 686–691.

Heimann, K., Umiltà, M. A., Guerra, M., & Gallese, V. (2014). Moving mirrors: a high-density EEG study investigating the effect of camera movements on motor cortex activation during action observation. *Journal of Cognitive Neuroscience*, *26*(9), 2087–2101.

Hogg, R. V., McKean, J. W., & Craig, A. T. (2013). *Introduction to mathematical statistics*. Pearson.

Hou, X., Harel, J., & Koch, C. (2012). Image signature: highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(1), 194–201.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Jasper, H. H. (1958). The ten-twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology*, *10*, 371–375.

Jenke, R., Peer, A., & Buss, M. (2014). Feature extraction and selection for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, *5*(3), 327–339.

John, V., Yoneda, K., Liu, Z., & Mita, S. (2015). Saliency map generation by the convolutional neural network for real-time traffic light detection using template matching. *IEEE Transactions on Computational Imaging*, *1*(3), 159–173.

Jolliffe, I. T. (1986). *Principal component analysis*. New York: Springer-Verlag.

Kajihara, T., Anwar, M. N., Kawasaki, M., Mizuno, Y., Nakazawa, K., & Kitajo, K. (2015). Neural dynamics in motor preparation: From phase-mediated global computation to amplitude-mediated local computation. *NeuroImage*, *118*, 445–455.

Kawasaki, M., Kitajo, K., & Yamguchi, Y. (2014). Fronto-parietal and fronto-temporal theta phase synchronization for visual and auditory-verbal working memory. *Frontiers in Psychology*, *5*(200), 1–7.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Klimesch, W., Schimke, H., & Schwaiger, J. (1994). Episodic and semantic memory: an analysis in the EEG theta and alpha band. *Electroencephalography and Clinical Neurophysiology*, *91*(6), 428–441.

Kloss, A., Kappler, D., Lensch, H. P. A., Butz, M. V., Schaal, S., & Bohg, J. (2016). Learning where to search using visual attention. In *2016 IEEE/RSJ international conference on intelligent robots and systems* (pp. 5238–5245).

Koch, C., & Ullman, S. (1985). Shift in selection in visual attention: toward the underlying neural circuitry. *Human Neurobiology*, *4*(4), 219–227.

Koivisto, M., Mäntylä, T., & Silvanto, J. (2010). The role of early visual cortex (V1/V2) in conscious and unconscious visual perception. *NeuroImage*, *51*(2), 828–834.

Lin, Y. P., Wang, C. H., Jung, T. P., Wu, T. L., Jeng, S. K., Duann, J. R., et al. (2010). EEG-based emotion recognition in music listening. *IEEE Transaction on Biomedical Engineering*, *57*(7), 1798–1806.

Mahmoodian, H., & Ebrahimian, L. (2016). Using support vector regression in gene selection and fuzzy rule generation for relapse time prediction of breast cancer. *Biocybernetics and Biomedical Engineering*, *36*(3), 466–472.

Mauthner, T., Possegger, H., Waltner, G., & Bischof, H. (2015). Encoding based saliency detection for videos and images. In *2015 IEEE conference on computer vision and pattern recognition* (pp. 2494–2502).

Michels, L., Moazami-Goudarzi, M., Jeanmonod, D., & Sarnthein, J. (2008). Eeg alpha distinguishes between cuneal and precuneal activation in working memory. *NeuroImage*, *40*(3), 1296–1310.

Murphy, K. R. (2012). *Machine learning: A probability perspective*. The MIT Press.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410.

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., & Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*, *63*(6), 902–915.

Neisser, U. (1967). Cognitive psychology, Appleton-Century-Crofts.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, *21*(19), 1641–1646.

Olmos, A., & Kingdom, F. A. (2004). A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception*, *33*, 1463–1473.

Parasuraman, R. (1998). *The attentive brain*. MIT Press.

Peng, H. C., Long, F., & Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(8), 1226–1238.

Ray, W. J., & Cole, H. W. (1985). EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes. *Science*, *228*, 750–752.

Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, *10*(5), 1299–1319.

Sharma, G., Jurie, F., & Schmid, C. (2012). Discriminative spatial saliency for image classification. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 3506–3513).

Stone, M. (1978). Cross-validation: a review. *Statistics*, *9*(1), 127–139.

Tavakoli, H. R., & Laaksonen, J. (2017). Bottom-up fixation prediction using unsupervised hierarchical models. In C. S. Chen, et al. (Eds.), *LNCS*: *Vol. 10116*. *ACCV 2016 workshops, Part I* (pp. 287–302).

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J. B., Lebihan, D., et al. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, *33*(4), 1104–1116.

Thomas, K. P., & Vinod, A. P. (2016). Utilizing individual alpha frequency and delta band power in EEG based biometric recognition. In *2016 IEEE international conference on systems, man, and cybernetics* (pp. 004787–004791).

Uribe, L. F. S., Fazanaro, F. I., Castellano, G., Suyama, R., Attux, R., Cardozo, E., et al. (2014). A recurrence-based approach for feature extraction in brain-computer interface systems. In *Book: Translational recurrences* (pp. 95–107). Springer International Publishing.

Welch, P. D. (1967). The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustic*, *15*(2), 70–73.

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision*, *14*(1), 1–20.

Yang, J., & Yang, M. H. (2017). Top-down visual saliency via joint CRF and dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(3), 576–588.

Zhang, J., & Sclaroff, S. (2013). Saliency detection: a Boolean map approach. In *2013 IEEE international conference on computer vision* (pp. 153–160).

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, *11*(3), 9, 1–15.

Zhong, S. H., Liu, Y., Liu, Y., & Chung, F. L. (2010). Asemantic no-reference image sharpness metric based on top-down and bottom-up saliency map modelling. In *2010 17th IEEE international conference on image processing* (pp. 1553–1556).

Zhou, X., Liu, Z., Sun, G., Ye, L., & Wang, X. (2016). Improving saliency detection via multiple kernel boosting and adaptive fusion. *IEEE Signal Processing Letters*, *23*(4), 517–521.